

Documentação da parte de Negócios do Projeto Big Data - Módulo 8 - Inteligência Artificial

Grupo Pérola Negra - Solução DataApp com Dashboard

Integrantes do Grupo:

- Ana Martire
- Eduardo Oliveira
- Keylla Oliveira
- Lucas Barbosa
- Nicollas Isaac
- Sophia Nóbrega

Sumário

- 1. Introdução
 - 1.1. Parceiro de Negócios
 - 1.2. Definição do Problema
 - 1.2.1. Problema
- 2. Objetivos
 - 2.1. Objetivos Gerais
 - 2.2. Objetivos Específicos
 - 2.3. Justificativa
- 3. Lean Inception
 - 3.1. Matriz É/Não é, Faz/Não faz
 - 3.2. Product Goal (Objetivos do Produto)
 - 3.3. Product Vision (Visão do Produto)
 - 3.4. Canvas MVP
- 4. Compreensão do Problema
 - 4.1. Canvas Proposta de Valor
 - 4.1.1. Perfil do Cliente
 - 4.1.2. Mapa de Valor
 - 4.2. Matriz de Risco
 - Sprint I
 - Sprint II
 - Sprint III
 - Sprint IV
 - Sprint V
 - Conclusão
 - 4.3. Total Addressable Market (TAM)
 - 4.4. Service Addressable Market (SAM)
 - 4.5. Service Obtainable Market (SOM)
- 5. Análise Financeira
 - 5.1. Custos de Implementação e Manutenção
 - 5.2. Custos de Desenvolvimento
 - 5.3. ROI (Return On Investment)
 - 5.4. Conclusão
- 6. Plano de Comunicação
 - 6.1. Objetivo
 - 6.2. Stakeholders
 - 6.3. Mensagens-Chave
 - 6.4. Canais de Comunicação
 - 6.5. Plano de Implementação
 - 6.6. Medidas de Sucesso
 - 6.7. Feedback e Ajustes
- 7. Conclusões
- 8. Referências

1. Introdução

Este projeto está sendo desenvolvido por alunos do quarto semestre do curso de Sistemas de Informação do Inteli, no âmbito do módulo 8 da graduação em 2024. A faculdade Inteli adota uma metodologia de aprendizado baseada em projetos (PBL - Problem-Based Learning), na qual os alunos aplicam o conhecimento teórico em situações práticas e reais. Este projeto, em particular, marca o encerramento do segundo ano de estudos dos alunos e envolve 34 estudantes. Sob a orientação do pós-doutor Renato Penha, e com o suporte de um corpo docente altamente qualificado, composto por professores com doutorado ou, no mínimo, mestrado, os alunos estão desafiados a criar uma solução real para um problema complexo.

1.1. Parceiro de Negócios

A Companhia Paulista de Trens Metropolitanos (CPTM) é uma sociedade de economia mista, criada pela Lei nº 7.861, de 28 de maio de 1992, sob a autorização do Poder Executivo do Estado de São Paulo. Regida pelo Artigo 158 da Constituição do Estado, a CPTM é responsável pela exploração dos serviços de transporte de passageiros sobre trilhos ou guiados nas regiões metropolitanas, aglomerações urbanas e microrregiões do Estado de São Paulo ([Alesp, 1996](#)). A empresa opera 57 estações, distribuídas em 5 linhas, que cobrem uma extensão de 196 quilômetros, transportando diariamente mais de 1,5 milhão de passageiros.

O papel da CPTM vai além do transporte de passageiros; ela tem uma importância estratégica para a mobilidade urbana e a qualidade de vida dos habitantes da maior metrópole do Brasil. Neste contexto, o projeto proposto para os alunos visa apoiar a empresa na análise de grandes volumes de dados operacionais, otimizando suas operações e contribuindo para a melhoria contínua dos seus serviços.

1.2. Definição do Problema

O grande volume de dados gerados pelas operações diárias da CPTM, provenientes de sistemas de controle e monitoramento, cria desafios consideráveis para a empresa, especialmente no que diz respeito à análise e interpretação dessas informações. A empresa enfrenta limitações tecnológicas e de infraestrutura que dificultam a extração de insights valiosos para a tomada de decisões estratégicas e operacionais.

1.2.1. Problema

O problema principal que o projeto aborda é a falta de recursos eficientes para analisar grandes volumes de dados gerados pelas operações da CPTM. A ausência de um pipeline de Big Data adequado para integrar, transformar e analisar esses dados impede a identificação de padrões relevantes e o uso de dados preditivos para otimizar as operações. Isso afeta diretamente a capacidade da empresa de melhorar a eficiência, reduzir custos e prever necessidades de manutenção e recursos operacionais.

2. Objetivos

Para a definição do projeto, é essencial estabelecer tanto os objetivos gerais quanto os específicos. A seguir, estão descritos os objetivos definidos para este projeto.

2.1. Objetivos Gerais

O objetivo geral deste projeto é desenvolver um pipeline de Big Data que permita à CPTM realizar análises estatísticas e descritivas em seus dados operacionais e administrativos. A solução proposta deverá ser capaz de lidar com grandes volumes de dados, proporcionando insights que melhorem a tomada de decisão da empresa, a gestão de recursos e a eficiência das suas operações. O projeto visa garantir que a CPTM possa explorar todo o potencial de seus dados, utilizando ferramentas modernas de processamento e análise em cloud, com foco em softwares de código aberto.

2.2. Objetivos Específicos

O projeto busca atingir objetivos específicos como:

1. Construir uma infraestrutura de Data Lake na AWS S3 para armazenamento eficiente de dados.
2. Desenvolver um sistema de ingestão de dados em batch e streaming, integrando dados de diversas fontes.
3. Implementar um processo de ETL (extração, transformação e carga) utilizando AWS Glue ou AWS Lambda para preparar os dados para análise.
4. Utilizar o EMR (Elastic MapReduce) com Apache Spark e Hadoop para análises estatísticas e descritivas.
5. Criar infográficos, utilizando AWS QuickSight ou ferramentas open-source, para a visualização dos resultados.
6. Garantir que a solução seja escalável e adaptável às futuras necessidades da CPTM.

2.3. Justificativa

A realização deste projeto é essencial para que a CPTM consiga superar suas limitações atuais na análise de dados e aproveite melhor os recursos tecnológicos disponíveis no mercado. A implementação de um pipeline de Big Data irá permitir à empresa otimizar seus processos operacionais, melhorar o planejamento estratégico e obter uma visão preditiva mais precisa, que pode resultar em redução de custos e aumento da eficiência. A adoção de tecnologias de cloud computing e Big Data também colocará a CPTM em um patamar mais competitivo, alinhado às melhores práticas de gestão de grandes volumes de dados em empresas de transporte público em todo o mundo.

3. Lean Inception

A metodologia **Lean Inception** é uma abordagem colaborativa usada para alinhar equipes na criação de produtos, combinando conceitos de **Design Thinking** e **Lean Startup** para definir e planejar o escopo do Produto Mínimo Viável (MVP). De acordo com Paulo Caroli, autor da metodologia, "Lean Inception é essencial para alinhar expectativas e criar um plano de ação claro para entregar valor incremental ao longo do desenvolvimento" (Caroli, 2024). Em projetos de grande escala como o da CPTM (Companhia Paulista de Trens Metropolitanos), onde há um alto volume de dados operacionais e múltiplos stakeholders, a Lean Inception torna-se um processo necessário para garantir que todos os envolvidos estejam em congruência, especialmente em um ambiente complexo de análise de dados e otimização de operações.

No contexto da CPTM, a Lean Inception foi utilizada para alinhar áreas chave, como o **Centro de Controle Operacional (CCO)**, a **gerência de manutenção**, e a **diretoria de operações**. O foco foi desenvolver uma solução de Big Data que centralize e analise dados operacionais e administrativos, permitindo otimizar o planejamento de manutenção e melhorar a eficiência operacional com base em dados. Ao aplicar a Lean Inception, foi possível estabelecer objetivos claros e delimitar o escopo do produto, utilizando ferramentas como a matriz "É/Não é, Faz/Não faz", Objetivo do produto, Visão do Produto e Canvas MVP.

3.1. Matriz É/Não é, Faz/Não faz

Uma das ferramentas centrais na Lean Inception é a matriz **É/Não é, Faz/Não faz**, que permite delimitar claramente o que o produto será e o que não será, além de listar as funcionalidades que ele entregará ou não. No caso do projeto da CPTM, essa matriz ajudou a esclarecer e alinhar as expectativas de todos os stakeholders, principalmente em relação à capacidade de processamento de dados e às funcionalidades esperadas para melhorar as operações da companhia. Veja ela a seguir:

Figura 1 - Matriz É/Não é, Faz/Não faz

É – Não É – Faz – Não Faz



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Com essa ferramenta, foi possível garantir que o pipeline de Big Data desenvolvido atenda às necessidades operacionais, sem promover funcionalidades fora do escopo, mantendo o foco em entregas práticas e relevantes.

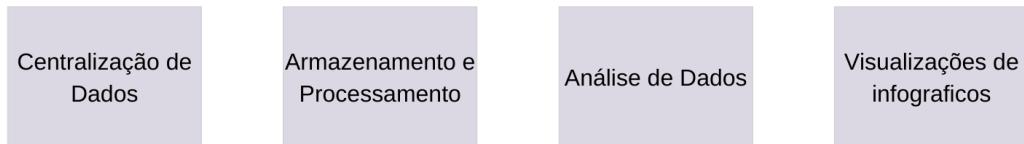
3.2. Product Goal (Objetivos do Produto)

O **Product Goal** define claramente o que o produto pretende alcançar. Ele serve como um guia central para o time de desenvolvimento e os stakeholders, garantindo que todos saibam quais resultados o produto deve entregar para ser considerado um sucesso. No projeto da CPTM, o objetivo principal é melhorar a eficiência operacional e o planejamento de manutenção por meio de uma solução centralizada de análise de dados.

O desenvolvimento desse pipeline de Big Data visa fornecer um sistema capaz de centralizar dados de diferentes áreas operacionais e administrativas, garantindo que as decisões estratégicas sejam fundamentadas em insights obtidos a partir de análises estatísticas descritivas. Esse alinhamento de dados permitirá à CPTM não apenas monitorar a eficiência das operações, mas também otimizar recursos, como o estoque de materiais consumíveis. Veja eles a seguir na figura:

Figura 2 - Product Goal

Product Goals



Fonte: Material produzido pelo Grupo Pélola Negra (2024)

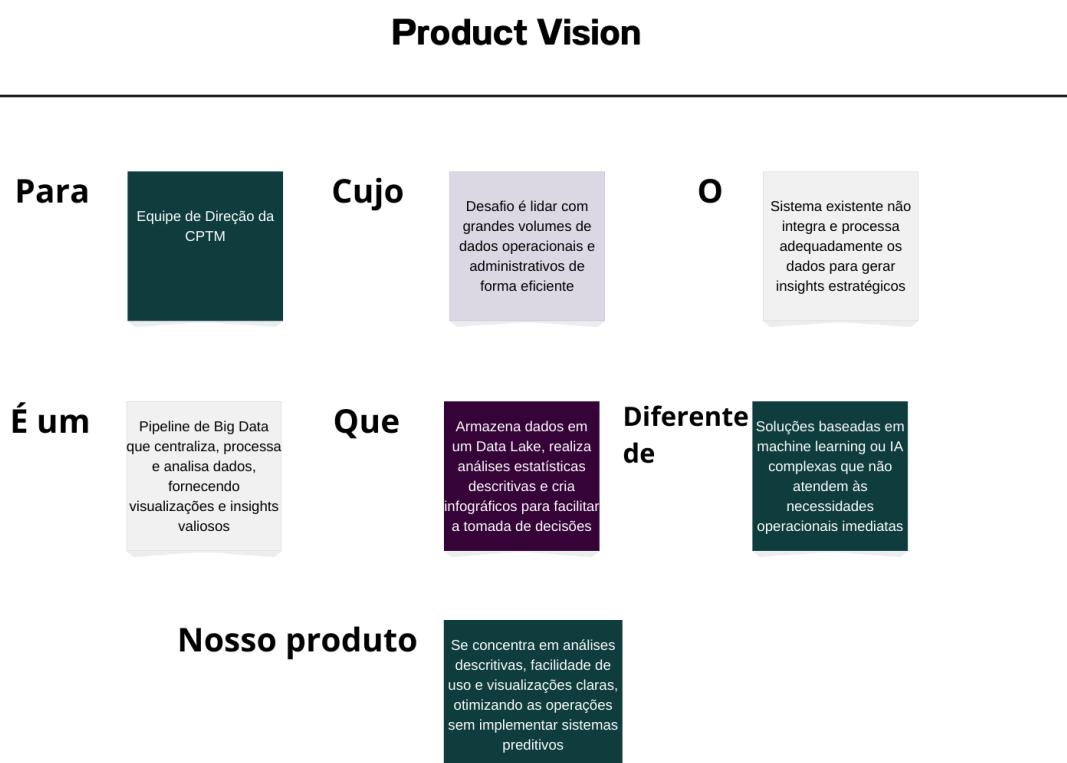
- Centralização de Dados:** Criar uma ferramenta única que centralize dados operacionais e administrativos em um ambiente seguro e acessível.
- Armazenamento e Processamento:** Proporcionar uma infraestrutura para armazenamento em grande escala de dados, com processamento eficiente por meio de tecnologias como AWS EMR e Apache Spark.
- Análise de Dados:** Executar análises estatísticas descritivas que ajudem a melhorar as operações e fornecer insights para o planejamento de manutenção.
- Visualizações e Relatórios:** Facilitar a tomada de decisões por meio de visualizações claras e relatórios detalhados criados com ferramentas como AWS QuickSight ou alternativas open-source.

3.3. Product Vision (Visão do Produto)

A **Product Vision** descreve a visão de futuro para o produto e como ele beneficiará os usuários finais. Ela fornece uma orientação clara e inspiradora para a equipe, assegurando que o produto entregue o máximo de valor possível. No caso da CPTM, o foco está em criar um pipeline de Big Data que possibilite o processamento e a análise centralizada de dados operacionais e administrativos, ajudando a empresa a tomar decisões mais embasadas.

A visão do produto para o projeto da CPTM é fornecer uma plataforma que centralize dados de diferentes áreas, processando-os de forma eficiente para gerar insights estratégicos que otimizem as operações diárias da empresa. A principal diferença do produto em comparação a outras soluções no mercado é seu foco em análises descritivas, ao invés de IA ou Machine Learning, o que torna o produto mais acessível, direto e fácil de implementar nas operações diárias da CPTM. Veja na figura 3 a seguir a tela usada para montar a visão do produto.

Figura 3 - Product Vision



Fonte: Material produzido pelo Grupo Pélola Negra (2024)

Ao focar em uma solução que centralize dados operacionais e administrativos e ofereça insights, a visão do produto reforça o compromisso de fornecer uma plataforma que seja acessível, eficiente e capaz de transformar o processo de tomada de decisões dentro da companhia.

3.4. Canvas MVP

O **Canvas MVP** é uma ferramenta que estrutura as principais componentes do MVP, como proposta de valor, segmentos de clientes, jornada do usuário e métricas, facilitando o planejamento e o acompanhamento do desenvolvimento. No projeto da CPTM, o Canvas MVP foi desenvolvido para garantir que o produto entregue soluções práticas e mensuráveis para os desafios operacionais identificados, como a centralização e análise de grandes volumes de dados. Veja ele a seguir:

Figura 4 - Canvas MVP



Fonte: Material produzido pelo Grupo Pélola Negra (2024)

O Canvas MVP estruturou o desenvolvimento do MVP de maneira que garantisse as funcionalidades prioritárias, os segmentos de clientes e as métricas alinhadas com os objetivos estratégicos da CPTM. Com o MVP bem definido, a equipe consegue direcionar esforços para entregar uma solução viável que gera valor imediato, facilitando o monitoramento das operações.

4. Compreensão do Problema

Esta seção se concentra na avaliação detalhada do ambiente e escopo geral do projeto, visando compreender as necessidades, aspirações e desafios enfrentados pelo parceiro, a CPTM. Além disso, busca-se examinar o desempenho atual do negócio, considerando a perspectiva do projeto em questão. Para isso, são coletados dados relevantes para o negócio, com o objetivo de identificar oportunidades de aprimoramento dentro do escopo do projeto e elaborar planos de ação estratégicos com uma visão de longo prazo para a implementação.

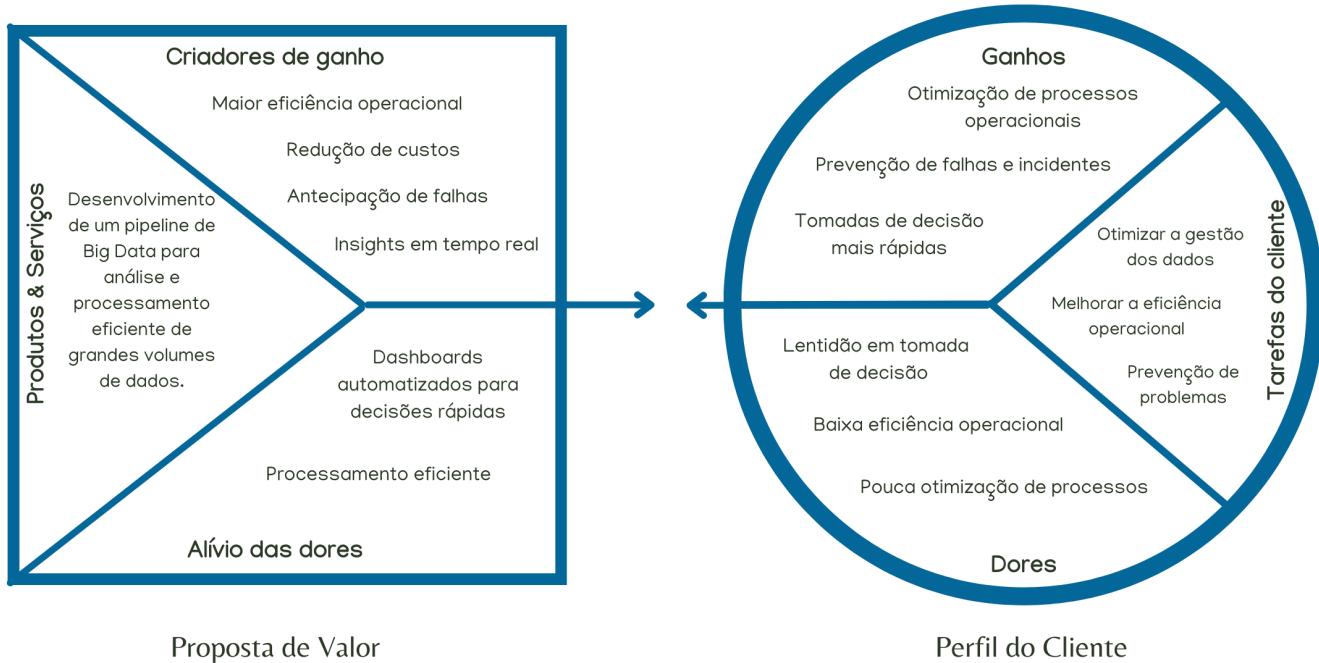
Dentro dessa análise, são considerados fatores como matriz de risco para identificar, avaliar e mitigar potenciais riscos que podem afetar negativamente o projeto e canvas de proposta de valor para compreender como o projeto se alinha aos objetivos e necessidades da CPTM. Essas ferramentas proporcionam uma compreensão abrangente do contexto empresarial, permitindo uma tomada de decisão informada e estratégica para o sucesso do projeto.

4.1. Canvas Proposta de Valor

De acordo com o grupo G4 educação: "O Value Proposition Canvas ou Proposta de Valor Canvas é uma ferramenta que permite aos empreendedores e empresários desenhar, testar e visualizar o valor do produto para os clientes, de uma forma intuitiva." Ao lado direito da imagem há o perfil do cliente, que apresenta os ganhos que ele terá com o produto, as dores atuais, e as tarefas dele no produto. Ao lado esquerdo há o mapa de valor, onde é apresentado o produto, seus criadores de ganho, e o alívio das dores do cliente. (Gushiken, 2024)

Figura 5 - Value Proposition Canvas

Value Proposition



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

4.1.1. Perfil do Cliente

Na análise do perfil do cliente, as principais tarefas esperadas do projeto são a **otimização da gestão dos dados**, a **melhoria da eficiência operacional** e a **prevenção de problemas**. A CPTM busca centralizar e organizar seus dados para facilitar o acesso e a tomada de decisões, além de automatizar processos repetitivos para melhorar a produtividade. Outro objetivo é implementar uma solução de monitoramento contínuo, capaz de antecipar falhas e evitar interrupções nos serviços.

O cliente tem algumas dores que quer que sejam sanadas com o projeto. Elas são a **lentidão nas tomadas de decisão**, a **baixa eficiência operacional** e a **pouca otimização de processos**. A falta de estrutura para processar e analisar grandes volumes de dados compromete a eficiência e gera atrasos na resposta a falhas. Além disso, a ausência de dados organizados dificulta decisões rápidas e informadas, enquanto a falta de uma visão consolidada prejudica a otimização coordenada das operações e a alocação eficiente de recursos.

Os "ganhos" demonstrados na figura do Value Proposition Canvas são o que a CPTM espera que aconteça com a solução que o grupo *Pérola Negra* desenvolveu. Eles estão diretamente ligados às "dores" descritas acima. Então, os ganhos que o cliente espera ter com o projeto são a **agilidade na tomada de decisões**, a **prevenção de falhas** e a **otimização de processos**. A CPTM espera que o projeto traga agilidade nas decisões, com dados acessíveis em tempo real, prevenção de falhas, através de monitoramento contínuo, e otimização de processos, eliminando gargalos e melhorando o uso de recursos. Esses ganhos "colidem" diretamente com as dores, ou seja, esses ganhos abordam de forma direta as principais dificuldades enfrentadas pela CPTM.

4.1.2. Mapa de Valor

Partindo agora para o mapa de valor, a parte esquerda da imagem, é possível notar que o produto que o grupo entregará à CPTM é o **desenvolvimento de um pipeline de Big Data para análise e processamento eficiente de grandes volumes de dados**. O pipeline visa centralizar e estruturar os dados da CPTM, proporcionando um processamento ágil e organizado. Essa solução oferece uma base sólida para automação de análises e geração de insights em tempo real, com uma estrutura escalável e adaptada para integrar diferentes fontes de dados. Além disso, facilita a criação de dashboards automatizados, fornecendo uma visão clara dos indicadores chave, apoiando a tomada de decisões e o monitoramento contínuo. Assim, o projeto não só entrega uma solução tecnológica, mas também uma ferramenta estratégica para otimizar a gestão dos dados e a eficiência operacional.

A seção de alívio das dores destaca os aspectos da solução que são projetados para resolver diretamente os desafios enfrentados pelo cliente. No caso da CPTM, a solução oferece **dashboards automatizados**, que permitem acesso rápido a informações críticas, proporcionando agilidade na tomada de decisões, e o **processamento eficiente dos dados**, que melhora a eficiência operacional, eliminando gargalos e reduzindo o tempo de resposta. Esses "pain killers", como são chamados em inglês, resolvem todas as dores que o cliente possui, as quais já foram citadas anteriormente.

Por fim, os criadores de ganho mostram o que a CPTM vai ganhar com os alívios das dores. Com a implementação do pipeline de Big Data, a empresa terá **maior eficiência operacional**, **reduzindo custos** e otimizando recursos através da automação de processos. Além disso, o acesso a **insights em tempo real** permitirá uma resposta mais rápida a incidentes e desafios operacionais, **antecipando possíveis falhas** e melhorando a tomada de decisões estratégicas.

O desenvolvimento de um Value Proposition Canvas é essencial para entender as necessidades do cliente e alinhar a solução com suas expectativas. Ele permite identificar de forma clara as principais dores e tarefas do cliente, os ganhos esperados, e como a solução proposta pode aliviá-las de maneira eficaz. Essa ferramenta estratégica ajuda a garantir que o produto ou serviço oferecido esteja diretamente conectado aos desafios reais enfrentados pelo cliente, fornecendo uma abordagem estruturada para criar valor. Além disso, facilita a comunicação da proposta entre a equipe e as partes interessadas, promovendo um entendimento comum e orientando o desenvolvimento de soluções que entreguem resultados concretos.

4.2. Matriz de Risco

A matriz de risco é uma ferramenta utilizada para a identificação, análise e priorização de riscos em projetos. Ela permite visualizar de forma clara e objetiva o impacto e a probabilidade de diferentes riscos, ajudando as equipes a focar naqueles que podem causar maiores prejuízos ou interrupções. Por meio de uma avaliação quantitativa ou qualitativa, a matriz de risco facilita a tomada de decisões, permitindo que medidas preventivas ou corretivas sejam implementadas de maneira rápida e eficaz.

Segundo Hillson, a matriz de risco é fundamental para a gestão proativa de riscos, uma vez que possibilita a priorização baseada em critérios claros e mensuráveis, como a probabilidade de ocorrência e a gravidade do impacto. Esta ferramenta ajuda a minimizar incertezas e a maximizar as chances de sucesso em projetos complexos, garantindo que os recursos sejam direcionados de forma eficiente para mitigar riscos críticos. (Hillson; Simon, 2024)

No projeto da Companhia Paulista de Trens Metropolitanos (CPTM), a matriz de risco é se faz necessária para gerenciar os desafios de uma iniciativa conduzida por 34 alunos do curso de Sistemas de Informação da Inteli. Com foco na criação de uma plataforma de dados para otimizar a gestão operacional e de manutenção, a ferramenta ajuda a identificar e priorizar riscos, como atrasos e falhas de comunicação com os stakeholders, além de aproveitar oportunidades, como feedbacks da CPTM e a colaboração técnica entre grupos. Essa abordagem garante que o projeto siga dentro do prazo e atenda às expectativas de todos stakeholders envolvidos.

Sprint I

Quadro 1 - Matriz de Risco 1

Probabilidade	Ameaças					Oportunidades				
	90%					Aprendizado de gestão de risco				
	70%		Prazo insuficiente	Falta de experiência técnica	Não compreensão dos dados informados	Possibilidade de otimização de processos	Colaboração intergrupal para conhecimento de bases			
	50%		Sobrecarga de trabalho	Perda de foco no MVP	Falta de alinhamento interno	Feedback valioso da CPTM	Fortalecimento da liderança	Capacitação interna técnica		
	30%		Retrabalho devido a decisões erradas	Falta da participação de membros na Daily.	Atraso na entrega das atividades.					
	10%			Comunicação ineficaz com stakeholders	Falta de suporte da CPTM					
	Muito Baixo	Baixo	Moderado	Alto	Muito Alto	Muito Alto	Alto	Moderado	Baixo	Muito Baixo
Impacto										
Sprint 1										

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Com a ciência da matriz de risco, agora é apresentado os responsáveis pelo gerenciamento dos riscos. A seguir, são detalhadas as oportunidades e ameaças identificadas, cada uma com sua respectiva probabilidade, impacto, responsável e plano de ação. Veja tudo a seguir:

Responsáveis no Projeto CPTM:

- **Time de Desenvolvimento:** Grupo Pérola Negra
- **Product Owner (PO):** Pessoas rotativas do grupo Pérola Negra
- **Scrum Master:** Pessoas rotativas do Grupo Pérola Negra
- **Orientador:** Renato Penha - Orientador da Turma 10 de Sistemas de Informação
- **Coordenador do Curso:** Egon - Coordenador do curso de Sistemas de Informação
- **Líder do Projeto:** Roberto Morina
- **Ponto Focal Backup:** Sarah de Sá Fernandes
- **Líder Técnico:** Roberto Morina
- **Líder de Negócio:** Sarah de Sá Fernandes
- **Líder Executivo:** Maicon Satiro de Oliveira

Oportunidades

1. Colaboração intergrupal para conhecimento de bases

- Impacto: Alto

- Probabilidade: 70%
- Descrição: Possibilidade de formar parcerias com outros grupos para compartilhar conhecimento técnico sobre as bases de dados.
- Responsável: Time de Desenvolvimento
- Plano de Ação: Organizar Daylies colaborativas e encontros regulares para troca de conhecimento entre os grupos.

2. Feedback valioso da CPTM

- Impacto: Muito Alto
- Probabilidade: 50%
- Descrição: A chance de receber insights detalhados da CPTM para melhorar o projeto.
- Responsável: Líder de Negócio e de Projeto
- Plano de Ação: Manter reuniões periódicas com stakeholders da CPTM para coletar feedback contínuo.

3. Otimização de processos

- Impacto: Muito Alto
- Probabilidade: 70%
- Descrição: Identificação de falhas no tratamento de dados pode gerar melhorias no processo.
- Responsável: Líder Técnico
- Plano de Ação: Implementar revisões semanais no processo de tratamento de dados para detectar e corrigir ineficiências.

4. Fortalecimento da liderança

- Impacto: Alto
- Probabilidade: 50%
- Descrição: Membros podem assumir papéis de liderança para organizar melhor o trabalho e fechar o escopo.
- Responsável: Scrum Master
- Plano de Ação: Alternar responsabilidades de liderança entre os membros do grupo para desenvolver habilidades de gestão.

5. Capacitação técnica interna

- Impacto: Moderado
- Probabilidade: 50%
- Descrição: Oportunidade para o grupo expandir seu conhecimento técnico.
- Responsável: Time de Desenvolvimento
- Plano de Ação: Agendar treinamentos técnicos em AWS, Apache Spark e Hadoop com foco nas ferramentas utilizadas no projeto.

6. Aprendizado em gestão de risco

- Impacto: Muito Alto
- Probabilidade: 90%
- Descrição: A gestão dos riscos neste projeto pode melhorar a capacidade do grupo de lidar com incertezas em futuros projetos.
- Responsável: Scrum Master
- Plano de Ação: Atualizar constantemente a matriz de risco e implementar boas práticas de gerenciamento de riscos.

Ameaças

1. Não compreensão dos dados fornecidos

- Impacto: Muito Alto
- Probabilidade: 70%
- Descrição: Dificuldade de interpretar corretamente os dados recebidos da CPTM pode atrasar o andamento do projeto.
- Responsável: Time de Desenvolvimento
- Plano de Ação: Realizar sessões de alinhamento com a equipe técnica da CPTM para garantir clareza nos dados.

2. Falta de participação nas dailies

- Impacto: Alto
- Probabilidade: 30%
- Descrição: Membros do grupo podem não participar ativamente das reuniões diárias, prejudicando o progresso.
- Responsável: Scrum Master
- Plano de Ação: Reforçar a importância das dailies e aplicar técnicas de gamificação para aumentar o engajamento.

3. Atraso na entrega das atividades

- Impacto: Muito Alto
- Probabilidade: 30%
- Descrição: Problemas de organização interna podem resultar em atrasos.
- Responsável: Product Owner
- Plano de Ação: Implementar o uso de ferramentas de gestão como o Jira para acompanhamento rigoroso das atividades, e se cabível negociar data de entrega com Orientador.

4. Falta de experiência técnica

- Impacto: Alto

- Probabilidade: 70%
- Descrição: O grupo pode enfrentar dificuldades técnicas para lidar com as ferramentas e dados.
- Responsável: Time de Desenvolvimento
- Plano de Ação: Identificar gaps de conhecimento e promover treinamentos adicionais com professores.

5. Retrabalho devido a decisões erradas

- Impacto: Moderado
- Probabilidade: 30%
- Descrição: Decisões incorretas no tratamento de dados podem resultar em retrabalho significativo.
- Responsável: Time de Desenvolvimento
- Plano de Ação: Revisar constantemente as decisões e ajustar os processos conforme o feedback das revisões.

6. Falta de alinhamento interno

- Impacto: Muito Alto
- Probabilidade: 50%
- Descrição: Desalinhamento entre membros pode afetar o desempenho geral.
- Responsável: Time de Desenvolvimento
- Plano de Ação: Realizar reuniões de alinhamento e refinamento interno semanalmente para discutir as responsabilidades de cada membro.

7. Prazo insuficiente

- Impacto: Moderado
- Probabilidade: 70%
- Descrição: O tempo necessário para concluir as soluções pode ser maior que o disponível.
- Responsável: Orientador
- Plano de Ação: Reavaliar constantemente o cronograma e ajustar as metas para garantir que as entregas essenciais sejam priorizadas de acordo com o plano de trabalho.

8. Falta de suporte da CPTM

- Impacto: Muito Alto
- Probabilidade: 10%
- Descrição: Caso a CPTM não forneça o suporte necessário, o projeto pode ser comprometido.
- Responsável: Líder do Projeto com Coordenador do Curso
- Plano de Ação: Estabelecer um canal de comunicação direto com a equipe da CPTM para garantir suporte contínuo.

9. Perda de foco no MVP

- Impacto: Alto
- Probabilidade: 50%
- Descrição: Falta de clareza sobre o objetivo principal pode levar a um MVP mal definido.
- Responsável: Coordenador de Escopo
- Plano de Ação: Realizar workshops para garantir que todos entendam claramente o foco do MVP e suas prioridades.

10. Sobrecarga de trabalho

- Impacto: Moderado
- Probabilidade: 50%
- Descrição: A equipe pode se sobrecarregar com as atividades.
- Responsável: Scrum Master
- Plano de Ação: Monitorar a carga de trabalho de cada membro e redistribuir tarefas quando necessário.

11. Comunicação ineficaz com stakeholders

- Impacto: Alto
- Probabilidade: 10%
- Descrição: Falhas de comunicação com stakeholders podem causar desalinhamento.
- Responsável: Product Owner
- Plano de Ação: Estabelecer um plano de comunicação claro com a CPTM e aproveitar datas de entrega para realizar reuniões de alinhamento.

Sprint II

Quadro 2 - Matriz de Risco 2

Probabilidade	Ameaças					Oportunidades					
	90%	70%	50%	30%	10%	Prazo insuficiente para o desenvolvimento	Falta de experiência técnica	Problemas de conectividade com Object Storage	Erros na transformação de dados	Capacitação em modelagem UML e Arquitetura	Possibilidade de otimização de processos
Muito Baixo	Baixo	Moderado	Alto	Muito Alto	Muito Alto						
Impacto											

Sprint 2

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Ameaças

1. Conflitos entre a equipe

- Impacto: Moderado
- Probabilidade: 30%
- Descrição: Desentendimentos entre membros da equipe podem afetar a produtividade e a união do time.
- Responsável: Scrum Master
- Plano de Ação: Promover sessões de feedback e conversas conjuntas para resolver conflitos de maneira rápida.

2. Problemas de conectividade com Object Storage

- Impacto: Alto
- Probabilidade: 50%
- Descrição: Falhas na conexão podem atrasar a extração de dados e interromper o ETL.
- Responsável: Time de Desenvolvimento
- Plano de Ação: Implementar soluções de backup de conexão e monitoramento contínuo.

3. Erros na transformação de dados

- Impacto: Muito Alto
- Probabilidade: 50%
- Descrição: Erros ao limpar e transformar dados podem resultar em informações inconsistentes para o carregamento no Data Warehouse.
- Responsável: Time de Desenvolvimento
- Plano de Ação: Implementar validação durante a etapa de transformação para garantir a integridade dos dados.

4. Prazo insuficiente para desenvolvimento

- Impacto: Moderado
- Probabilidade: 70%
- Descrição: Falta de tempo adequado para concluir todas as tarefas pode levar a um trabalho apressado e de menor qualidade.
- Responsável: PO
- Plano de Ação: Priorizar tarefas críticas e ajustar o escopo durante a daily, se necessário.

5. Falta de experiência técnica

- Impacto: Alto
- Probabilidade: 70%
- Descrição: A equipe pode não ter experiência e conhecimento suficiente em algumas tecnologias, o que pode atrasar o trabalho e comprometer a qualidade do resultado.
- Responsável: Time de Desenvolvimento
- Plano de Ação: Identificar áreas de deficiência técnica e providenciar o suporte e estudo necessários.

Oportunidades

1. Possibilidade de otimização de processos

- Impacto: Alto
- Probabilidade: 50%
- Descrição: Melhorias na etapa de transformação podem resultar em um processo mais escalável.
- Responsável: Time de Desenvolvimento

- Plano de Ação: Avaliar regularmente o desempenho do processo e implementar melhorias caso necessário.

2. Capacitação técnica interna

- Impacto: Moderado
- Probabilidade: 30%
- Descrição: O projeto oferece a oportunidade para que a equipe desenvolva novas habilidades técnicas, especialmente em ETL.
- Responsável: Scrum Master e Equipe
- Plano de Ação: Realizar os autoestudos e estar presente nas aulas, buscando atendimento individual com o professor, se necessário.

3. Capacitação em modelagem UML e Arquitetura

- Impacto: Alto
- Probabilidade: 70%
- Descrição: O projeto oferece à equipe a chance de aprimorar suas habilidades em modelagem UML e Arquitetura, o que pode ser muito útil em projetos que exijam a documentação de arquiteturas de sistemas.
- Responsável: Time de Desenvolvimento
- Plano de Ação: Realizar os autoestudos específicos em UML e práticas de modelagem para garantir a precisão e qualidade do diagrama.

Sprint III

Quadro 3 - Matriz de Risco 3

		Ameaças				Oportunidades					
Probabilidade	Impacto	90%									
		70%		Prazo insuficiente para o desenvolvimento	Falta de experiência técnica	Problemas com tecnologias novas (Prefect)		Aprendizagem com relação a visualização dos dados			
		50%		Falta da presença do grupo completo em momentos importantes	Problemas de conectividade com Object Storage	Erros na transformação de dados			Melhoria na convivência e trabalho em equipe do grupo		
		30%		Conflitos entre a equipe					Capacitação técnica interna		
		10%									
	Muito Baixo		Baixo	Moderado	Alto	Muito Alto	Muito Alto	Alto	Moderado	Baixo	Muito Baixo
Sprint 3		Impacto									

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Ameaças

1. Prazo insuficiente para o desenvolvimento

- Impacto: Muito Alto
- Probabilidade: 90%
- Descrição: O curto prazo pode prejudicar a entrega ou qualidade das entregas.
- Responsável: Product Owner
- Plano de Ação: Priorizar o planejamento e a realização de tarefas sem atrasos..

2. Falta de experiência técnica

- Impacto: Alto
- Probabilidade: 90%
- Descrição: A falta de conhecimento técnico pode prejudicar as entregas.
- Responsável: Time de Desenvolvimento
- Plano de Ação: Realizar estudos além dos propostos em sala para entender as tecnologias e fluxos que serão utilizados.

3. Problemas de conectividade com Object Storage

- Impacto: Muito Alto
- Probabilidade: 50%
- Descrição: Dificuldades técnicas de conectividade podem atrasar as entregas.
- Responsável: Time de Desenvolvimento
- Plano de Ação: Fazer testes constantes com a conectividade, e sempre alinhar com os professores.

4. Conflitos entre a equipe

- Impacto: Moderado

- Probabilidade: 30%
- Descrição: Falta de alinhamento pode gerar conflitos e afetar as entregas.
- Responsável: Scrum Master
- Plano de Ação: Durante os encontros ajudar a mediar as discussões.

5. Falta da presença do grupo completo em momentos importantes

- **Impacto:** Moderado
- **Probabilidade:** 50%
- **Descrição:** A ausência de membros durante os encontros prejudica o grupo no alinhamento e no planejamento das tarefas.
- **Responsável:** Scrum Master
- **Plano de Ação:** Garantir o comprometimento do grupo com os encontros diários e marcar a presença dos membros.

6. Problemas com tecnologias novas (Prefect)

- **Impacto:** Muito Alto
- **Probabilidade:** 70%
- **Descrição:** A introdução de uma tecnologia nova no projeto, pode dificultar e muito o desenvolvimento da entrega em especial quando se trata de uma tecnologia como Prefect, que não se encontra um número elevado de documentos/materiais para ajuda e estudo.
- **Responsável:** Time de Desenvolvimento
- **Plano de Ação:** Buscar conhecer sobre a tecnologia e entender quais serão os usos dentro do projeto.

7. Erros na transformação de dados

- **Impacto:** Muito Alto
- **Probabilidade:** 50%
- **Descrição:** Falhas no processo de transformação dos dados podem gerar um trabalho não planejado, e um atraso nas entregas.
- **Responsável:** Time de Desenvolvimento
- **Plano de Ação:** Planejar todas as entregas com antecedência, e sempre alinhar as dificuldades com o grupo.

Oportunidades

1. Aprendizagem com relação à visualização dos dados

- Impacto: Moderado
- Probabilidade: 90%
- Descrição: Oportunidade de entender melhor sobre as técnicas de visualização para análise e apresentação dos dados, utilizando tecnologias novas como streamlit.
- Responsável: Time de Desenvolvimento
- Plano de Ação: Estudar ferramentas de visualização.

2. Melhoria na convivência e trabalho em equipe do grupo

- Impacto: Alto
- Probabilidade: 50%
- Descrição: Possibilidade de alinhar a equipe e fortalecer a colaboração entendendo melhor os erros e acertos das sprints anteriores.
- Responsável: Scrum Master
- Plano de Ação: Fazer sessões de feedback e dinâmicas de grupo para melhorar a comunicação.

3. Capacitação técnica interna

- Impacto: Moderado
- Probabilidade: 50%
- Descrição: Aumento do conhecimento técnico do grupo.
- Responsável: Time de Desenvolvimento
- Plano de Ação: Estudar além das aulas e autoestudos as tecnologias que serão necessárias.

Sprint IV

Quadro 4 - Matriz de Risco 4

Probabilidade	Ameaças					Oportunidades				
	90%						Aprendizagem com relação a novas tecnologias (Streamlit)			
	70%		Prazo insuficiente para o desenvolvimento	Falta de experiência técnica			Aprendizagem com relação a visualização dos dados			
	50%			Falta da presença do grupo completo em momentos importantes	Problemas devido a refatoração e retrabalho		Possibilidade de otimização de processos	Melhoria na convivência e trabalho em equipe do grupo		
	30%		Conflitos entre a equipe		Falta de comprometimento em relação às entregas e atividades			Capacitação técnica interna		
	10%									
	Muito Baixo	Baixo	Moderado	Alto	Muito Alto	Muito Alto	Alto	Moderado	Baixo	Muito Baixo
Sprint 4										
Impacto										

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Ameaças

1. Prazo insuficiente para o desenvolvimento

- Impacto: Muito Alto
- Probabilidade: 90%
- Descrição: O curto prazo pode prejudicar a entrega ou qualidade das entregas.
- Responsável: Product Owner
- Plano de Ação: Priorizar o planejamento e a realização de tarefas sem atrasos.

2. Falta de experiência técnica

- Impacto: Alto
- Probabilidade: 90%
- Descrição: A falta de conhecimento técnico pode prejudicar as entregas.
- Responsável: Time de Desenvolvimento
- Plano de Ação: Realizar estudos além dos propostos em sala para entender as tecnologias e fluxos que serão utilizados.

3. Problemas devido à refatoração e retrabalho

- Impacto: Muito Alto
- Probabilidade: 50%
- Descrição: Falta de comprometimento no planejamento pode ocasionar em situações de diversos membros fazendo itens semelhantes e deixando de se dedicar a outros itens importantes.
- Responsável: Product Owner
- Plano de Ação: Realizar os encontros com a equipe comprometida, e alinhar sempre todas as tarefas sendo realizadas.

4. Falta de comprometimento em relação às entregas e atividades

- Impacto: Alto
- Probabilidade: 30%
- Descrição: Baixo comprometimento de membros pode prejudicar as entregas.
- Responsável: Scrum Master
- Plano de Ação: Acompanhar de perto o progresso e sempre alinhar as tarefas nos encontros do grupo.

5. Conflitos entre a equipe

- Impacto: Moderado
- Probabilidade: 30%
- Descrição: Falta de alinhamento pode gerar conflitos e afetar as entregas.
- Responsável: Scrum Master
- Plano de Ação: Durante os encontros ajudar a mediar as discussões.

6. Falta da presença do grupo completo em momentos importantes

- **Impacto:** Moderado
- **Probabilidade:** 50%
- **Descrição:** A ausência de membros durante os encontros prejudica o grupo no alinhamento e no planejamento das tarefas.
- **Responsável:** Scrum Master
- **Plano de Ação:** Garantir o comprometimento do grupo com os encontros diários e marcar a presença dos membros.

Oportunidades

1. Aprendizagem com relação a novas tecnologias (Streamlit)

- Impacto: Muito Alto
- Probabilidade: 70%
- Descrição: Oportunidade de adotar e aprender novas tecnologias úteis para o projeto.
- Responsável: Time de Desenvolvimento
- Plano de Ação: Buscar entender e aprender durante a realização de tarefas de novas tecnologias.

2. Aprendizagem com relação à visualização dos dados

- Impacto: Moderado
- Probabilidade: 90%
- Descrição: Oportunidade de entender melhor sobre as técnicas de visualização para análise e apresentação dos dados, utilizando tecnologias novas como streamlit.
- Responsável: Time de Desenvolvimento
- Plano de Ação: Estudar ferramentas de visualização.

3. Possibilidade de otimização de processos

- Impacto: Alto
- Probabilidade: 50%
- Descrição: Melhorias no fluxo de trabalho identificando falhas e diminuindo retrabalhos.
- Responsável: Líder Técnico
- Plano de Ação: Entender as ultimas sprints e as tecnologias, para que os fluxos possam ser otimizados.

4. Melhoria na convivência e trabalho em equipe do grupo

- Impacto: Alto
- Probabilidade: 50%
- Descrição: Possibilidade de alinhar a equipe e fortalecer a colaboração entendendo melhor os erros e acertos das sprints anteriores.
- Responsável: Scrum Master
- Plano de Ação: Fazer sessões de feedback e dinâmicas de grupo para melhorar a comunicação.

5. Capacitação técnica interna

- Impacto: Moderado
- Probabilidade: 50%
- Descrição: Aumento do conhecimento técnico do grupo.
- Responsável: Time de Desenvolvimento
- Plano de Ação: Estudar além das aulas e autoestudos as tecnologias que serão necessárias.

Sprint V

Quadro 5 - Matriz de Risco 5

		Ameaças					Oportunidades				
Probabilidade	90%					Tempo insuficiente para a entrega final	Melhoria no DataAPP com base em feedback				
	70%				Sobrecarga de trabalho		DataAPP interativo	Entrega do Pipeline			
	50%			Falta de alinhamento interno							
	30%										
	10%										
		Muito Baixo	Baixo	Moderado	Alto	Muito Alto	Muito Alto	Alto	Moderado	Baixo	Muito Baixo
	Sprint 5	Impacto									

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Ameaças

1. Tempo insuficiente para a entrega final

- Impacto: Muito Alto

- Probabilidade: 90%
- Descrição: Apenas uma semana de trabalho pode comprometer a qualidade e o prazo das entregas.
- Responsável: Scrum Master
- Plano de Ação: Priorizar as tarefas mais críticas e realizar revisões diárias para mitigar atrasos.

2. Sobrecarga de Trabalho

- Impacto: Alto
- Probabilidade: 70%
- Descrição: A pressão para cumprir o prazo pode sobrecarregar os membros da equipe, afetando a produtividade.
- Responsável: PO
- Plano de Ação: Dividir as tarefas de forma equilibrada e estabelecer períodos curtos de descanso.

3. Falta de alinhamento interno

- Impacto: Moderado
- Probabilidade: 50%
- Descrição: Divergências entre os membros sobre prioridades podem atrasar o progresso das entregas.
- Responsável: Scrum Master
- Plano de Ação: Promover reuniões rápidas de alinhamento e priorizar a comunicação clara.

Oportunidades

1. Melhoria no DataAPP com base em feedback

- Impacto: Muito Alto
- Probabilidade: 90%
- Descrição: Incorporar melhorias com base nos feedbacks pode aumentar significativamente a qualidade do produto.
- Responsável: Time de Desenvolvimento
- Plano de Ação: Priorizar o feedback crítico e implementá-lo rapidamente.

2. Entrega do Pipeline

- Impacto: Alto
- Probabilidade: 70%
- Descrição: Demonstrar a arquitetura completa do pipeline pode reforçar a competência técnica do grupo.
- Responsável: Time de Desenvolvimento
- Plano de Ação: Garantir que todas as etapas do pipeline sejam devidamente integradas e testadas.

3. DataAPP interativo

- Impacto: Muito Alto
- Probabilidade: 70%
- Descrição: Entregar um aplicativo com alto nível de interatividade pode agregar valor ao projeto e garantir alta usabilidade.
- Responsável: Time de Desenvolvimento
- Plano de Ação: Realizar testes e ajustar funcionalidades com base nos resultados.

Conclusão

A matriz de risco, acompanhada dos planos de ação, facilita o reconhecimento de oportunidades que podem preparar e fortalecer o grupo, ao mesmo tempo em que permite uma abordagem preventiva diante das ameaças identificadas. Com esse mapeamento claro, a equipe está mais bem equipada para evitar situações que possam prejudicar o andamento do projeto ou afetar stakeholders importantes. Assim, o grupo consegue atuar de forma mais estratégica, garantindo tanto a mitigação de riscos quanto o aproveitamento de oportunidades para o sucesso do MVP.

4.3. Total Addressable Market (TAM)

A Companhia Paulista de Trens Metropolitanos (CPTM) desempenha um papel fundamental no transporte público do Estado de São Paulo, servindo uma vasta população que depende desse sistema ferroviário para seus deslocamentos diários. Para entender o potencial econômico deste serviço, é essencial analisar três componentes-chave: o **Mercado Total Endereçável (TAM)**, o **Mercado Disponível e Endereçável (SAM)** e o **Mercado Obtível (SOM)**. Esses parâmetros permitem avaliar o alcance e a viabilidade financeira da CPTM dentro de uma das maiores regiões metropolitanas do Brasil.

Como a CPTM é uma empresa pública do Governo de São Paulo, seus clientes em potencial são todos os cidadãos do estado. No entanto, para calcular o **TAM**, é preciso considerar o alcance geográfico dos serviços oferecidos. Assim, o **TAM** da CPTM inclui a população dos municípios por onde suas linhas passam:

1. **Caieiras**: 95.032
2. **Campo Limpo Paulista**: 77.632
3. **Ferraz de Vasconcelos**: 179.198
4. **Francisco Morato**: 165.139
5. **Franco da Rocha**: 144.849
6. **Guarulhos**: 1.291.771
7. **Itaquaquecetuba**: 369.275
8. **Jundiaí**: 443.221

9. **Mauá:** 418.261
10. **Mogi das Cruzes:** 451.505
11. **Poá:** 103.765
12. **Ribeirão Pires:** 115.559
13. **Rio Grande da Serra:** 44.170
14. **Santo André:** 748.919
15. **São Caetano do Sul:** 165.655
16. **São Paulo:** 11.451.999
17. **Suzano:** 307.429
18. **Várzea Paulista:** 115.771

Com uma população total de **16.689.150 pessoas** nos municípios atendidos, o **TAM** reflete o potencial máximo de mercado em termos de bilhetes diários. Considerando um valor de R\$5,00 por bilhete e a hipótese de que cada pessoa faça duas viagens por dia (ida e volta), o **TAM** diário seria:

$$\text{TAM} = 16.689.150 \text{ pessoas} \times \text{R\$5,00} \times 2 \text{ bilhetes} = \text{R\$166.891.500,00} \text{ em bilhetes diários potenciais.}$$

4.4. Service Addressable Market (SAM)

Para calcular o **SAM**, é necessário excluir a população que tem acesso a transporte particular, como carros ou motos. Segundo o IPEA, **54% dos domicílios no Sudeste possuem um veículo**, o que reduz a necessidade de transporte público para uma parcela significativa da população.

Portanto, **46%** da população das cidades atendidas pela CPTM ainda depende do transporte público. Logo, o **SAM** pode ser calculado como:

$$\text{SAM} = 46\% \text{ de } 16.689.150 \text{ pessoas} = 7.677.009 \text{ pessoas} \text{ (sem acesso regular a carro/moto).}$$

O potencial de receita em bilhetes diários seria:

$$\text{SAM} = 7.677.009 \text{ pessoas} \times \text{R\$5,00} \times 2 \text{ bilhetes} = \text{R\$76.770.090,00} \text{ em bilhetes diários potenciais.}$$

4.5. Service Obtainable Market (SOM)

Para calcular o **SOM**, utilizamos dados reais sobre o número de passageiros diários da CPTM. Atualmente, cerca de **3,2 milhões de pessoas** utilizam o sistema de trens diariamente.

Logo, o **SOM** em termos de receita diária seria:

$$\text{SOM} = 3.200.000 \text{ pessoas} \times \text{R\$5,00} \times 2 \text{ bilhetes} = \text{R\$32.000.000,00} \text{ em bilhetes diários potenciais.}$$

A análise do **TAM**, **SAM** e **SOM** permite uma visão clara do potencial de mercado da CPTM. Com um **TAM** de R\$166,9 milhões em bilhetes diários, reduzido para um **SAM** de R\$76,7 milhões, e um **SOM** mais realista de R\$32 milhões, é possível ver o impacto direto dos fatores de mobilidade e posse de veículos na receita potencial. Essas métricas auxiliam no planejamento e na tomada de decisões para otimizar o serviço da CPTM e aumentar sua eficiência operacional e financeira.

5. Análise Financeira

A análise financeira é uma etapa essencial em qualquer processo de avaliação e tomada de decisão relacionado a novos investimentos. Essa prática abrange diferentes tipos de organizações, como empresas privadas, entidades sem fins lucrativos e instituições estatais, garantindo uma avaliação aprofundada da viabilidade e dos riscos envolvidos, além de fornecer bases sólidas para decisões estratégicas.

5.1. Custos de Implementação e Manutenção

Na análise financeira para a implementação e manutenção do projeto, são considerados aspectos fundamentais, como a utilização de premissas predefinidas e a necessidade de contratação de mão de obra especializada. Esses fatores são cruciais para assegurar a precisão das projeções e o sucesso na execução.

Abaixo, é apresentada uma tabela com a estrutura de custos esperada para a implementação e manutenção do projeto, considerando que ele será realizado *on premise* (executado em software local).

Tabela 1 - Estrutura de Custos

Categoria	Nome	Custo Fracionado	Custo Total	Fonte
Infraestrutura	Servidor Local	R\$0,00 (Estado)	R\$0,00	CPTM
Tecnologia	Docker	R\$0,00 (Open Source)	R\$0,00	Próprio Site
Tecnologia	DBeaver	R\$0,00 (Open Source)	R\$0,00	Próprio Site
Tecnologia	Python	R\$0,00 (Open Source)	R\$0,00	Próprio Site
Tecnologia	Prefect	R\$0,00 (Open Source)	R\$0,00	Próprio Site
Tecnologia	Streamlit	R\$0,00 (Open Source)	R\$0,00	Próprio Site
Mão de Obra	2 Cientistas de Dados	R\$30,77/hora cada	R\$8.100,20 (cada)	Talent, 2024

Categoría	Nome	Custo Fracionado	Custo Total	Fonte
Mão de Obra	2 Analistas de Dados	R\$23,28/hora cada	R\$6.458,40 (cada)	Talent, 2024
Mão de Obra	Engenheiro de Software	R\$41,54/hora	R\$10.460,99	Talent, 2024
Mão de Obra	Product Owner	R\$36,92/hora	R\$9.448,28	Talent, 2024
Mão de Obra	Product Manager	R\$12,31/hora	R\$4.053,77	Talent, 2024
TOTAL/MÊS	-	-	R\$53.080,24/mês	-
TOTAL/ANO	-	-	R\$636.962,93/ano	-

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Detalhamento dos Custos por Membro da Equipe

A tabela abaixo apresenta os custos detalhados de cada cargo, incluindo salário base, benefícios e encargos tributários, para fornecer uma visão completa dos valores investidos na equipe.

Tabela 2 - Estrutura de Custos Detalhados

Cargo	Escopo	Valor/Hora	Carga Horária	Salário Total	Benefícios (VC S/A, 2024)	Renda Total	Impostos (37% do salário líquido) (Catho, 2024)	Custo Total
2 Cientistas de Dados	Otimização do ETL	R\$30,77/hora	160h/mês	R\$4.923,20	VR: R\$1.135,42VT: R\$220,00	R\$6.278,62	R\$1.821,58	R\$8.100,20 (cada)
2 Analistas de Dados	Criação de Views e Dashboards	R\$23,28/hora	160h/mês	R\$3.724,80	VR: R\$1.135,42VT: R\$220,00	R\$5.080,22	R\$1.378,18	R\$6.458,40 (cada)
Engenheiro de Software	Front-end e Arquitetura	R\$41,54/hora	160h/mês	R\$6.646,40	VR: R\$1.135,42VT: R\$220,00	R\$8.001,82	R\$2.459,17	R\$10.460,99
Product Owner	Gestão do projeto	R\$36,92/hora	160h/mês	R\$5.907,20	VR: R\$1.135,42VT: R\$220,00	R\$7.262,62	R\$2.185,66	R\$9.448,28
Product Manager	Recolhimento de demandas e métricas	R\$12,31/hora	160h/mês	R\$1.969,60	VR: R\$1.135,42VT: R\$220,00	R\$3.325,02	R\$728,75	R\$4.053,77
TOTAL/MÊS	-	-	-	R\$31.819,20/mês	-	R\$41.307,14/mês	R\$11.773,10/mês	R\$53.080,24/mês
TOTAL/ANO	-	-	-	-	-	-	-	R\$636.962,93/ano

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Explicações Importantes

Aqui, é relevante ressaltar a diferença entre **salário** e **renda**:

- **Salário:** Refere-se ao montante associado às horas trabalhadas.
- **Renda:** Inclui o total recebido, considerando benefícios e outras fontes de receita.

Os conceitos acima foram detalhados para demonstrar os custos totais de cada funcionário, desde o salário base até os impostos que a empresa deve arcar sobre cada cargo.

5.2. Custos de Desenvolvimento

Além dos custos de implementação e manutenção, também é apresentada a estrutura de custos relacionada à ideação e ao desenvolvimento do projeto. Isso garante transparência e detalhamento sobre os recursos utilizados, sejam tecnologias, infraestrutura ou mão de obra.

Tabela 3 - Estrutura de Custos de Desenvolvimento

Categoría	Nome	Custo Fracionado	Custo Total
Infraestrutura	Amazon S3	R\$0,00 (Patrocínio)	R\$0,00
Tecnologia	Docker	R\$0,00 (Open Source)	R\$0,00
Tecnologia	DBeaver	R\$0,00 (Open Source)	R\$0,00
Tecnologia	Python	R\$0,00 (Open Source)	R\$0,00

Categoria	Nome	Custo Fracionado	Custo Total
Tecnologia	Prefect	R\$0,00 (Open Source)	R\$0,00
Tecnologia	Streamlit	R\$0,00 (Open Source)	R\$0,00
Mão de Obra	Desenvolvedores (6 alunos)	R\$0,00 (Alunos)	R\$0,00
TOTAL	-	-	R\$0,00

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Observação

Embora seja possível calcular um valor representativo para os custos de desenvolvimento, ele não reflete a realidade, pois, na prática, não houve aporte financeiro da cliente (CPTM). Assim, optou-se por descrever os recursos utilizados para manter a transparência, sem atribuir custos monetários fictícios.

5.3. ROI (Return On Investment)

Com base nos custos detalhados, é possível estimar as receitas potenciais que o projeto pode gerar para a CPTM. De acordo com a análise TAM-SAM-SOM, cerca de 3,2 milhões de pessoas utilizam transporte público diariamente, enquanto o público endereçável da CPTM é estimado em aproximadamente 7,7 milhões de indivíduos. Melhorias no sistema ferroviário têm demonstrado resultados significativos em outros países. Na Europa, por exemplo, houve aumentos expressivos no uso de trens após reformas: na Espanha, o incremento foi de 26%, e na Itália, de 18,7% ([Railway, 2023](#)).

Esses dados sugerem que melhorias no sistema ferroviário têm o potencial de aumentar significativamente a utilização dos serviços e, consequentemente, as receitas geradas. Aplicando uma estimativa conservadora de aumento de 10% no uso diário dos trens da CPTM, isso representaria uma receita adicional de R\$3,2 milhões por dia. Em um mês, considerando apenas dias úteis, o incremento seria de R\$70,4 milhões, totalizando R\$844,8 milhões em um ano.

Com base nesses números, o ROI (Retorno sobre Investimento) do projeto pode ser calculado. O ROI é uma métrica essencial para avaliar a viabilidade de um investimento, sendo obtido pela fórmula: subtraímos o custo total do investimento da receita adicional gerada, dividimos o resultado pelo custo total e multiplicamos por 100. No caso deste projeto, com um custo total estimado de R\$636.962,928, o ROI é calculado como:

$$\text{ROI} = [(70.400.000 - 636.962,928) / 636.962,928] \times 100 = 109,52 \times 100 = 10.952,44\%$$

Este resultado reflete um retorno excepcionalmente elevado, evidenciando o enorme potencial do projeto para gerar impacto financeiro positivo para a CPTM, além de reforçar a viabilidade e a atratividade do investimento.

5.4. Conclusão

A análise financeira realizada evidencia a importância de um planejamento detalhado e transparente para o sucesso do projeto. Com a identificação clara dos custos de implementação, manutenção e desenvolvimento, foi possível estabelecer uma base sólida para avaliar a viabilidade do investimento e suas implicações estratégicas.

Além disso, a projeção de possíveis retornos com base em estudos e casos internacionais demonstra que melhorias no sistema ferroviário têm o potencial de gerar impactos positivos, tanto no aumento da utilização do transporte quanto nas receitas da CPTM. Esses indicadores reforçam a relevância do projeto e o papel essencial da gestão eficiente de recursos para alcançar resultados sustentáveis.

Portanto, a decisão de avançar com o projeto deve considerar tanto os benefícios tangíveis quanto os intangíveis, garantindo que as melhorias propostas atendam às expectativas de usuários e stakeholders, contribuindo para o desenvolvimento da infraestrutura de transporte público e para a qualidade de vida na região.

6. Plano de Comunicação

Esta seção detalha o plano de comunicação estabelecido para o projeto com a CPTM, incluindo objetivos, stakeholders, mensagens-chave, canais de comunicação, plano de implementação, medidas de sucesso, feedback e ajustes. O objetivo é garantir uma comunicação eficaz entre todos os envolvidos, promovendo alinhamento, transparéncia e eficiência.

6.1. Objetivo

O plano de comunicação tem como objetivo assegurar que todos os stakeholders estejam alinhados em relação ao progresso, desafios e entregas do projeto. Ele promove transparéncia, colaboração e uma tomada de decisão eficiente, facilitando o fluxo de informações entre os diferentes níveis do projeto. Isso minimiza riscos e maximiza a eficiência da equipe.

6.2. Stakeholders

Esta seção apresenta os stakeholders do projeto, descrevendo seus papéis e como se comunicam entre si. Também explicita a hierarquia e responsabilidades.

Principais Stakeholders e seus Papéis

- **Time de Desenvolvimento:**
 - **Grupo Pérola Negra:** Responsável pelo desenvolvimento técnico da solução.

- **Product Owner (PO):**

- **Membros Rotativos do Grupo Pérola Negra:** Coordenam as necessidades do cliente com a equipe de desenvolvimento.

- **Quadro de PO:**

Quadro 6 - Quadro de PO

Sprint Product Owner		
----- -----		
Sprint 1 Nicolas		
Sprint 2 Lucas		
Sprint 3 Eduardo		
Sprint 4 Ana		
Sprint 5 Nicollas		

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

- **Scrum Master:**

- **Membros Rotativos do Grupo Pérola Negra:** Facilitam a metodologia ágil e garantem a remoção de impedimentos.

- **Quadro de Scrum Master:**

Quadro 7 - Quadro de Scrum Master

Sprint Scrum Master		
----- -----		
Sprint 1 Sophia		
Sprint 2 Keylla		
Sprint 3 Sophia		
Sprint 4 Eduardo		
Sprint 5 Lucas		

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

- **Orientador:**

- **Renato Penha:** Orientador da Turma 10 de Sistemas de Informação, fornece direcionamento técnico e metodológico.

- **Coordenador do Curso:**

- **Egon Daxbacher:** Coordena o curso e oferece suporte acadêmico ao projeto.

- **Líder do Projeto:**

- **Roberto Morina:** Gerencia a equipe e assegura que as metas e entregas sejam alcançadas.

- **Ponto Focal Backup:**

- **Sarah de Sá Fernandes:** Serve como ponto de contato secundário, garantindo continuidade em casos de ausência do líder.

- **Líder Técnico:**

- **Roberto Morina:** Fornece liderança técnica e orienta o desenvolvimento da solução.

- **Líder de Negócio:**

- **Sarah de Sá Fernandes:** Garante que a solução atenda aos objetivos de negócios e à estratégia da CPTM.

- **Líder Executivo:**

- **Maicon Satiro de Oliveira:** Representa a alta gestão da CPTM e toma decisões estratégicas para o projeto.

Comunicação entre Stakeholders:

- O **Time de Desenvolvimento** reporta diretamente ao **Scrum Master** e ao **Product Owner**, que organizam e priorizam as entregas.
- O **Scrum Master** facilita a comunicação com o orientador e a alta gestão.
- O **Product Owner** é responsável por comunicar os requisitos recebidos da CPTM para o time de desenvolvimento.
- A **Sarah de Sá Fernandes**, como ponto focal backup e líder de negócio, intermedia possíveis pedidos do time de desenvolvimento ao time técnico da CPTM.
- O **Orientador** revisa entregas e garante a aderência às diretrizes do curso.

Para facilitar o acesso às informações sobre os stakeholders, veja a tabela abaixo, que traz o stakeholder, se ele é interno ou externo, suas expectativas e sua influência:

Tabela 4 - Expectativa e Influência

Stakeholder	Tipo	Expectativa	Influência
Time de Desenvolvimento	Interno	Concluir o desenvolvimento técnico da solução de forma eficiente, atendendo aos requisitos e prazos estabelecidos.	Alta - Responsável pela entrega técnica e resolução de desafios.
Product Owner (PO)	Interno	Garantir que os requisitos da CPTM sejam entendidos e priorizados corretamente, comunicando as necessidades do cliente ao time de desenvolvimento.	Alta - Define prioridades e reflete as expectativas do cliente no produto.
Scrum Master	Interno	Facilitar a execução da metodologia ágil, removendo impedimentos e garantindo o progresso contínuo do projeto.	Média - Influencia o andamento das sprints, mas não define requisitos diretamente.
Orientador (Renato Penha)	Interno	Fornecer direcionamento técnico e metodológico, ajudando a manter o projeto dentro dos padrões acadêmicos e profissionais exigidos.	Alta - Atua como guia técnico e garante alinhamento com a metodologia acadêmica.
Coordenador do Curso	Interno	Assegurar que o projeto esteja alinhado aos objetivos do curso e às expectativas da instituição de ensino.	Média - Garante alinhamento acadêmico, mas não interfere diretamente no desenvolvimento diário.
Líder do Projeto	Externo	Gerenciar o time, alinhar entregas e coordenar as interações entre stakeholders internos e externos.	Alta - Centraliza a organização e garante a entrega dos objetivos do projeto.
Ponto Focal Backup	Externo	Assegurar continuidade em casos de ausência do líder, respondendo por questões críticas e mantendo alinhamento com os stakeholders principais.	Média - Atua como suporte, mas sua influência é limitada à ausência do líder.
Líder Técnico	Externo	Fornecer suporte técnico ao time de desenvolvimento e assegurar a qualidade técnica das entregas.	Alta - Central na supervisão técnica e qualidade das soluções implementadas.
Líder de Negócio	Externo	Garantir que a solução atenda às metas estratégicas da CPTM e aos objetivos do cliente, validando requisitos de negócio.	Alta - Influencia diretamente na adequação do projeto às expectativas estratégicas da CPTM.
Líder Executivo (Maicon)	Externo	Representar a alta gestão da CPTM, assegurando que o projeto esteja alinhado aos objetivos estratégicos e aprovando decisões críticas.	Alta - Principal tomador de decisões estratégicas e aprovador final das entregas do projeto.
CPTM (Gestão e Operação)	Externo	Receber atualizações claras sobre o progresso do projeto e entender como ele impactará a operação, esperando uma solução prática e viável.	Alta - Define os requisitos gerais e valida a usabilidade da solução proposta.
Time Técnico da CPTM	Externo	Garantir que os requisitos técnicos estejam claros e sejam cumpridos, fornecendo suporte técnico quando necessário.	Média - Influencia tecnicamente o escopo, mas depende de alinhamentos internos para mudanças estruturais.
Faculdade Inteli	Interno	Observar o desempenho do grupo no projeto, avaliando como ele reflete o aprendizado e os objetivos do curso, além de manter a parceria com a CPTM.	Média - Não interfere diretamente no projeto, mas influencia o alinhamento geral com os objetivos educacionais.

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

E para tornar o fluxo de comunicação entre os stakeholders mais intuitivo, veja a figura abaixo, que apresenta um diagrama de comunicação entre eles, ilustrando tudo que foi falado anteriormente. Outras informações são aplicados nas demais subseções da seção 6.

Figura 6 - Diagrama de Stakeholders



Ferramentas de Comunicação



Meet

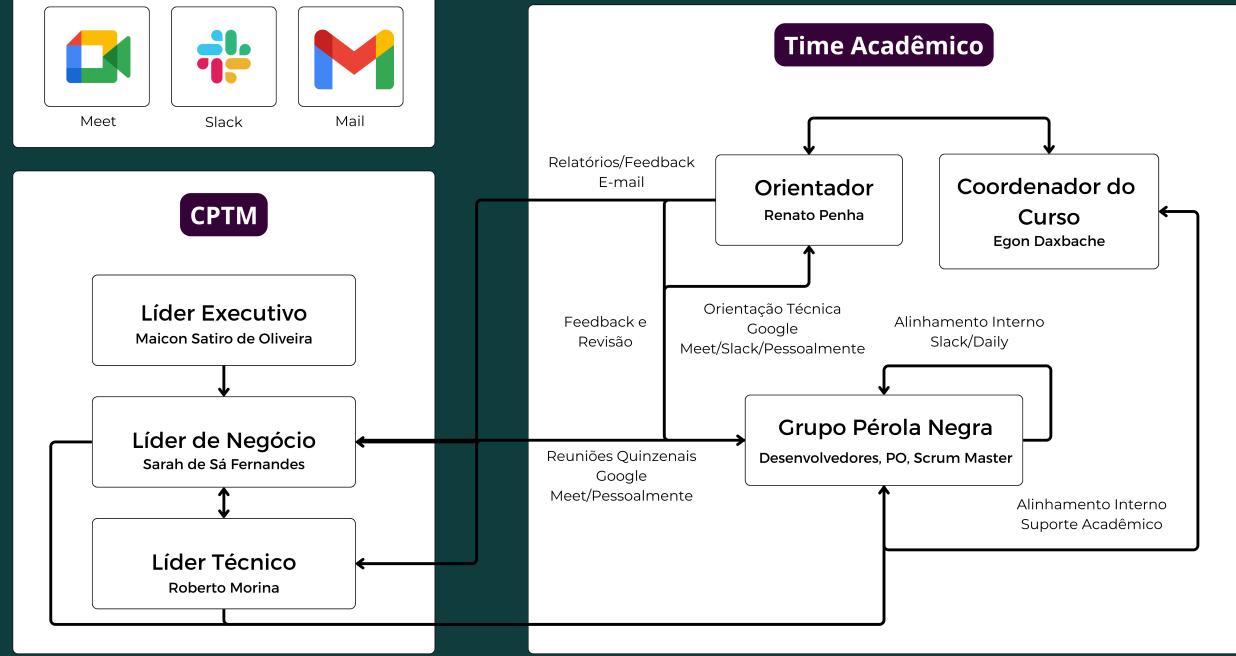


Slack



Mail

Diagrama de comunicação entre Stakeholders



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

6.3. Mensagens-Chave

Esta seção define as principais mensagens que serão transmitidas aos stakeholders, adaptadas às suas necessidades e expectativas.

Principais Mensagens:

- Para a CPTM (Gestão e Operação):
 - Atualizações sobre o progresso do projeto.
 - Alinhamento sobre requisitos e impactos operacionais.
- Para o Time Acadêmico:
 - Orientações sobre as entregas técnicas.
 - Suporte metodológico e acadêmico.
- Para a Equipe de Desenvolvimento:
 - Priorização de tarefas.
 - Refinamento de funcionalidades.
 - Resolução de impedimentos.

6.4. Canais de Comunicação

Nesta seção, descrevemos os canais de comunicação escolhidos, considerando a natureza de cada mensagem e a frequência ideal de uso.

- Reuniões:
 - Daily Meetings (Grupo Pérola Negra): Realizadas ao longo da semana para refinamento de tarefas e alinhamento interno.
 - Reuniões de Entrega de Sprint: Realizadas quinzenalmente com Sarah de Sá Fernandes para discutir entregas e alinhar dúvidas.
- Ferramentas Digitais:
 - Google Meet: Para comunicação entre stakeholders e reuniões virtuais.
 - E-mail: Para formalizações e envio de relatórios e comunicação emergencial.
 - Slack: Para comunicação rápida entre os membros do Grupo Pérola Negra.

6.5. Plano de Implementação

Esta seção descreve como o plano de comunicação será implementado, detalhando etapas e responsabilidades.

Etapas do Plano:**1. Definição da Frequência:**

- Dailies semanais para alinhamento técnico.
- Reuniões quinzenais com os principais stakeholders (Sarah e Roberto Morina).

2. Responsáveis pela Comunicação:

- O Scrum Master e o Product Owner garantem que o plano seja seguido e que todos sejam atualizados, sendo eles que apresentam as atividades feitas ao longo de cada sprint.

3. Controle de Qualidade:

- Revisão das mensagens e feedback constante dos stakeholders para garantir clareza e eficiência.

6.6. Medidas de Sucesso

Os indicadores de sucesso avaliados para o plano de comunicação incluem:

- Taxa mínima de participação nas reuniões: **90%**.
- Tempo médio para resolução de dúvidas pendentes: **inferior a 72 horas**.
- Feedback positivo dos stakeholders em relação à clareza e eficiência da comunicação: **70% ou mais**.
- Completude de mais de **70% das tarefas** da sprint, apresentadas no status report na entrega da sprint.

6.7. Feedback e Ajustes

Esta seção explica como o feedback será coletado e os ajustes realizados.

Plano para Ajustes:**1. Coleta de Feedback:**

- Questionários quinzenais para avaliar a efetividade dos canais, mensagens e entregas da sprint.
- Sessões de revisão com o orientador (Renato Penha) para alinhamento com o TAPI (Termo de Abertura do Projeto do Inteli) e Lean Inception da seção 3.

2. Ajustes Necessários:

- Redefinição da frequência das reuniões, se necessário.
- Adaptação dos canais de comunicação para atender às demandas do projeto.

7. Conclusões

Mais do que um simples resumo do que já foi dito, as conclusões do projeto mostram a evolução de toda a equipe e a construção de um alicerce sólido para que a CPTM passe a tomar decisões guiadas por dados. Não foi só alinhar objetivos, delimitar escopo ou fazer análises estatísticas, mas também avanços para uma visão integrada, que combina a mudança cultural na forma de lidar com informação e a criação de uma base tecnológica robusta, segura e escalável.

Partimos de um diagnóstico realista, olhando para as brechas na operação e para o potencial inexplorado do volume de dados gerados diariamente. Assim, desenhamos um pipeline de Big Data não para ter mais números, mas sim para transformá-los em insights práticos e úteis. Além das ferramentas técnicas usadas (como Data Lake, ETL e dashboards), buscamos criar um projeto para, futuramente, aplicar previsões mais precisas, análises em tempo real e um aperfeiçoamento constante das rotinas de manutenção e planejamento.

Porém, não ficamos só na parte técnica. Ao esclarecer quem são os stakeholders, como se dá a comunicação entre eles e como manter um fluxo de feedback constante, ampliamos o impacto da solução. Isso garante que o conhecimento gerado não se perca na complexidade da empresa. Ao integrar áreas internas da CPTM e conectar o time técnico ao gerencial e acadêmico, a iniciativa deixa de ser um produto fechado para se tornar um processo contínuo de melhoria. Cada sprint, ajuste ou análise alimenta um ciclo que faz a empresa crescer em maturidade.

A análise financeira reforça a importância estratégica de apostar em dados. Não é só economizar ou aumentar a eficiência: é colocar a CPTM em outro patamar competitivo, algo especialmente relevante num serviço público que influencia a mobilidade e a qualidade de vida da população.

Chegar até aqui não significou apenas colocar um projeto no papel, mas criar as bases para um ecossistema de dados mais esperto. O próximo passo não é simplesmente "codificar o que falta", mas consolidar essas diretrizes como um padrão de qualidade. O objetivo é que, a partir desse ponto, cada novo insight ajude a CPTM a antecipar demandas, melhorar o serviço, economizar recursos e, principalmente, aprender continuamente.

8. Referências

ALESP. Lei nº 7.861, de 28 de maio de 1992. 28 maio 1996. Disponível em: <https://www.al.sp.gov.br/repositorio/legislacao/lei/1992/lei-7861-28.05.1992.html>. Acesso em: 19 out. 2024.

CAROLI, Paulo. 3 Differences Between Design Sprint and Lean Inception You Need To Know. Disponível em: <https://caroli.org/3-differences-between-design-sprint-and-lean-inception-you-need-to-know/>. Acesso em: 20 ago. 2024.

CAROLI, Paulo. Lean Inception: How to Align People and Build the Right Product. Disponível em: <https://caroli.org/lean-inception-3/>. Acesso em: 20 ago. 2024.

CAROLI, Paulo. Learn Lean Inception at Caroli.org. Disponível em: <https://caroli.org/lean-inception-how-to-align-people-and-build-the-right-product/>. Acesso em: 20 ago. 2024.

CAROLI, Paulo. Why did the Lean Inception creator get involved with Data Mesh?. Disponível em: <https://caroli.org/why-did-the-lean-inception-creator-get-involved-with-data-mesh/>. Acesso em: 20 ago. 2024.

CARVALHO, Leandro S. Data Product Canvas. Disponível em: <https://medium.com/@leandrosarvalho/data-product-canvas-cd91f24776b1>. Acesso em: 10 set. 2024.

GUSHIKEN, A. (2023, 23 de outubro). Value Proposition Canvas: o que é e como funciona essa metodologia? G4 Educação. <https://g4educacao.com/portal/value-proposition-canvas>. Acesso em: 17 out. 2024.

HILLSON, D.; SIMON, P. Practical project risk management : the ATOM methodology. Tysons Corner, Va.: Management Concepts, 2012.

STICKDORN, M.; SCHNEIDER, J. This is service design thinking basics, tools, cases. [s.l.] Amsterdam Bis Publ, 2015.

TALENT.COM. (2024). Salário médio de Cientista de Dados em Brasil 2024. Talent.com. <https://br.talent.com/salary?job=cientista+de+dados#:~:text=Sal%C3%A1rio%20M%C3%A9dio%20de%20Cientista%20De%20Dados%20em%20Brasil%202024&text=O%20sal%C3%A1rio%20m%C3%A9dio%20de%20cientista,a%20ganhar%20R%24105.003%20anuais>. Acesso em: 3 dez. 2024.

TALENT.COM. (2024). Salário médio de Analista de Dados em Brasil 2024. Talent.com. <https://br.talent.com/salary?job=analista+de+dados#:~:text=Sal%C3%A1rio%20M%C3%A9dio%20de%20Analista%20De%20Dados%20em%20Brasil%202024&text=O%20sal%C3%A1rio%20m%C3%A9dio%20de%20analista,a%20ganhar%20R%2474.755%20anuais>. Acesso em: 3 dez. 2024.

TALENT.COM. (2024). Salário médio de Engenheiro de Software. Talent.com. <https://br.talent.com/salary?job=engenheiro+de+software>. Acesso em: 3 dez. 2024.

TALENT.COM. (2024). Salário médio de Product Owner. Talent.com. <https://br.talent.com/salary?job=product+owner>. Acesso em: 3 dez. 2024.

TALENT.COM. (2024). Salário médio de Product Manager. Talent.com. <https://br.talent.com/salary?job=product+manager>. Acesso em: 3 dez. 2024.

RAILWAY SUPPLY. (2023). Significant growth in rail passenger transport across Europe in 2023. Railway Supply. <https://www.railway.supply/en/significant-growth-in-rail-passenger-transport-across-europe-in-2023/>. Acesso em: 3 dez. 2024.

EUROSTAT. (2023). Railway passenger transport statistics - quarterly and annual data. Eurostat. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Railway_passenger_transport_statistics_-_quarterly_and_annual_data. Acesso em: 3 dez. 2024.

VOCÊ S/A. Vale-refeição dura 10 dias por mês no Brasil, mostra pesquisa. Disponível em: <https://vocesa.abril.com.br/economia/vale-refeicao-dura-10-dias-por-mes-no-brasil-mostra-pesquisa>. Acesso em: 8 dez. 2024.

CATHO PARA EMPRESAS. Quanto custa um funcionário para a empresa? Disponível em: <https://paraempresas.catho.com.br/quanto-custa-um-funcionario-para-empresa/#:~:text=Somados%2C%20os%20encargos%20equivalem%20a,da%20empresa%20vai%20precisar%20desembolsar>. Acesso em: 8 dez. 2024.

Documentação da parte de Experiência do Usuário do Projeto Big Data - Módulo 8 - Inteli

Grupo Pérola Negra - Solução DataApp com Dashboard

Integrantes do Grupo:

- [Ana Martire](#)
- [Eduardo Oliveira](#)
- [Keylla Oliveira](#)
- [Lucas Barbosa](#)
- [Nicollas Isaac](#)
- [Sophia Nóbrega](#)

Sumário

- [1. Análise de Experiência do Usuário](#)
 - [1.1. Personas](#)
 - [1.2. Jornada do Usuário](#)
 - [1.3. User Stories](#)
 - [1.4. Wireframe](#)
- [2. Conclusões](#)
- [3. Referências](#)

1. Análise de Experiência do Usuário

1.1. Personas

No desenvolvimento de sistemas e soluções, compreender o público-alvo é crucial, por isso, as personas são uma ferramenta importante oferecendo uma visão detalhada das necessidades, comportamentos e objetivos dos usuários. Essas representações fictícias, mas baseadas em pesquisa, permitem aos designers e equipes de desenvolvimento visualizar melhor quem serão os usuários finais do produto ou sistema, facilitando a criação de soluções que realmente atendam às suas expectativas e necessidades. A seguir, está a apresentado a persona formulada para o projeto:

Figura 1 - Imagem Representativa da Persona



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Sérgio Ribeiro, é um engenheiro de produção formado pela Escola politécnica da Universidade de São Paulo, ele atua como gestor operacional da CPTM (Companhia Paulista de Trens Metropolitanos), onde trabalha há 25 anos. Com 52 anos, Sérgio tem muita experiência no setor de transporte ferroviário e ele ficou conhecido por sua habilidade em melhorar a área operacional da companhia, deixando ela mais eficiente. Durante a carreira, ele geriu diversas iniciativas para otimizar o transporte dos passageiros na região paulista, tentando manter um equilíbrio saudável entre os custos e a qualidade do transporte e da experiência do usuário. Ele é casado, tem dois filhos, e nas horas vagas gosta de estudar, ouvir música, e praticar piano.

No momento, o Sérgio enfrenta muitos desafios, mas em especial a falta de ferramentas adequadas para a análise de grandes volumes de dados. Ele deseja aumentar a produtividade e a qualidade para a análise de dados, para também ajudar a melhorar a qualidade dos serviços da CPTM. Ele enxerga que uma boa solução de Big Data pode ser a solução para esses problemas, o que ajudaria em uma gestão mais eficiente das operações e na melhora da experiência dos passageiros. Sérgio acredita que esta solução poderia marcar ainda mais o nome da CPTM como uma referência no setor de transporte.

Sérgio participa das reuniões de planejamento estratégico da empresa, onde ele ajuda os líderes técnicos e executivos na hora de alinhar as diretrizes deste projeto de Big Data que está sendo desenvolvido. Ele também acompanha de perto a implementação das novas tecnologias nas operações diárias, que vão ajudar no andamento do projeto. O Sérgio acredita que a tecnologia é o futuro, e que projetos assim vão ter um

papel crucial na gestão dos dados da CPTM, e ele se dedica a explorar novas formas de utilizar a análise de dados para prevenir falhas operacionais.

O objetivo do Sérgio é garantir que a empresa continue evoluindo e que consiga manter uma boa qualidade de serviço. Ele acredita que "a eficiência operacional é fundamental, e a análise de dados pode nos ajudar a prever falhas antes que elas aconteçam." Para Sérgio, o futuro do transporte metropolitano está completamente ligado à inovação tecnológica, e ele está comprometido em liderar essa transformação dentro da CPTM. Para a visualização da persona, veja a figura a seguir:

Figura 2 - Detalhes sobre a Persona



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Para ajudar na análise da persona desenvolvida, o grupo também desenvolveu um mapa de empatia desta persona, explicando o que a persona pensa, sente, vê, ouve, faz e também quais são suas dores e ganhos, a seguir está a imagem que foi desenvolvida para o mapa de empatia:

Figura 3 - Mapa de Empatia da Persona



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Na seção do mapa "O que ele pensa e sente", é possível compreender diversas nuances das emoções e motivações do Sérgio. Sua preocupação surge da pressão por resultados positivos, mas a falta de uma análise de dados precisa gera frustração. Ao mesmo tempo, ele encontra esperança na possibilidade de um futuro mais promissor, impulsionado por ferramentas mais avançadas. Por fim, a determinação de Sérgio é alimentada por sua paixão pelo trabalho e o desejo constante de aperfeiçoar processos e alcançar uma performance mais eficiente.

Com essa persona, é possível reforçar como fazer essa tarefa de personificação ajuda a direcionar as decisões de design com base em uma compreensão profunda do comportamento e das necessidades do usuário, evitando armadilhas comuns como o design auto-referente e o usuário elástico. Além disso, elas facilitam a comunicação dos achados da pesquisa entre os membros da equipe, garantindo que todos tenham uma compreensão clara dos usuários alvo. ([Babich, 2024](#))

A Interaction Design Foundation descreve o Design Centrado no Usuário (UCD) como um processo iterativo que se concentra nas necessidades e nos requisitos dos usuários em cada etapa do processo de design. Através das etapas de pesquisa, definição de requisitos, design e avaliação, as equipes podem garantir que os produtos finais sejam úteis e utilizáveis para as pessoas. O UCD encoraja a inclusão dos usuários no processo de design, utilizando uma variedade de técnicas de pesquisa e design para criar produtos altamente acessíveis e usáveis. ([IxDF, 2024](#))

1.2. Jornada do Usuário

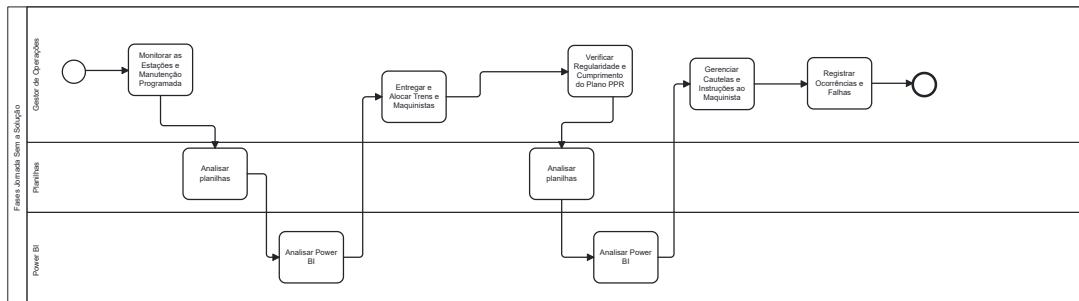
A jornada do usuário é uma ferramenta fundamental no design centrado no usuário, pois permite mapear a experiência completa de um usuário ao interagir com um produto ou sistema. Ela detalha os passos que o usuário percorre, os pontos de contato e as emoções em cada fase, identificando oportunidades de melhoria e pontos de fricção.

Este mapeamento ajuda a equipe de design e desenvolvimento a entender melhor o comportamento dos usuários em diferentes cenários, garantindo que a solução seja intuitiva e eficiente. A jornada do usuário é composta por várias etapas que refletem as interações com o sistema, desde o momento em que o usuário descobre o produto até a realização de seus objetivos.

De acordo com Stickdorn & Schneider (2011), uma jornada de usuário bem estruturada não apenas melhora a usabilidade do sistema, mas também cria um entendimento mais profundo das necessidades e motivações dos usuários, contribuindo para um design mais eficiente e centrado nas reais expectativas do público-alvo. ([Stickdorn; Schneider, 2024](#))

Abaixo está as fases da jornada da persona, Sérgio Ribeiro, sem a solução do Grupo Pérola Negra:

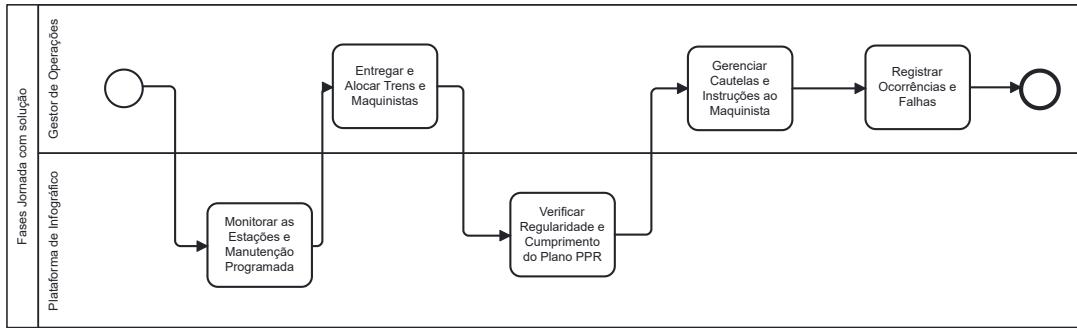
Figura 4 - BPMN sem a Solução



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Abaixo está as fases da jornada da persona, Sérgio Ribeiro, com a solução do Grupo Pérola Negra:

Figura 5 - BPMN com a Solução



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Com a nossa solução, os processos de gestão são mais otimizados e os dados ficam centralizados na plataforma do Grupo Pérola Negra, sendo assim, planilhas e Power BI não são necessários.

Com base na persona, segue o mapa de jornada do usuário, para uma melhor visualização segue o link: https://miro.com/app/board/uXjVLRSKuSE=/?share_link_id=775458606363

1.3. User Stories

Nesta seção, serão explorados os principais aspectos que guiam o desenvolvimento de uma solução centrada no usuário. A análise começa com a criação de personas, representações detalhadas dos usuários finais que permitem uma compreensão profunda de suas necessidades, expectativas e desafios. Em seguida, será apresentada a jornada do usuário, mapeando os principais pontos de interação com o sistema, desde o primeiro contato até a conclusão de suas metas.

As user stories são outra parte essencial do projeto que serão abordadas, que descrevem de forma clara as funcionalidades que o sistema precisa oferecer para atender às necessidades dos usuários. Por fim, serão incluídos wireframes, representações visuais que auxiliam na comunicação do fluxo e da interface da solução proposta, garantindo que o design atenda aos objetivos do usuário.

De acordo com Ricardo Arruda, uma User Story é uma representação clara e informal que expressa as necessidades e/ou requisitos de um potencial usuário, sendo considerada uma parte fundamental para atingir um objetivo final. Essa abordagem é a menor unidade de trabalho, centrada na necessidade do cliente final que irá interagir com o produto. ([Arruda, 2024](#))

No contexto do desenvolvimento de software, as User Stories são uma técnica ágil amplamente adotada para descrever funcionalidades sob a perspectiva do usuário. Estruturadas na forma "**Como** (quem), **quero** (o quê), **para** (finalidade)", essas histórias simplificam a comunicação entre desenvolvedores e stakeholders, priorizando as necessidades reais do usuário.

A seguir é possível visualizar 5 User Stories criadas para o contexto desse projeto, sendo condizentes com a persona e sua jornada utilizando a solução.

Tabela 1 - User Story 1 - Gerenciar Cautelas e Instruções

Número	1
Título	Gerenciar Cautelas e Instruções
Descrição	<p>Como gestor de operações, eu quero gerenciar as cautelas e instruções dadas aos maquinistas, para assegurar que eles recebam as informações corretas para operar com segurança.</p>
Critérios de Aceitação	<ol style="list-style-type: none"> 1. O sistema deve permitir o envio automático de cautelas e instruções com registro de confirmação de recebimento por parte dos maquinistas. 2. As instruções devem incluir informações como rota, horários e alertas de segurança.
Testes de Aceitação	<p>Critério de Aceitação 01: Cautelas enviadas automaticamente com registro de recebimento.</p> <ul style="list-style-type: none"> - Registro disponível e confirmado = correto. - Registro ausente ou incompleto = incorreto, precisa de ajuste. <p>Critério de Aceitação 02: Instruções completas visíveis no sistema.</p> <ul style="list-style-type: none"> - Instruções exibidas corretamente = correto. - Instruções ausentes ou incorretas = incorreto, precisa de ajuste.
Prioridade	Alta (Must Have). Garantir que os maquinistas tenham acesso às informações críticas de segurança é essencial para evitar incidentes e manter a operação segura.

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

A User Story 1 atende ao modelo INVEST. Ela é **independente**, abrangendo somente o gerenciamento de cautelas e instruções para os maquinistas, sem dependência de outras funcionalidades. A história é **negociável**, podendo ajustar o formato e a frequência de envio das instruções, conforme necessário. É **valiosa**, pois garante que o gestor forneça informações críticas de segurança de forma organizada e eficiente, ajudando a manter as operações seguras. A User Story é **estimável** com clareza de escopo, o que facilita a previsão de esforço para implementação. É também **pequena**, focando apenas na organização e envio de informações de cautela, e **testável**, com critérios definidos para verificar a clareza e o recebimento adequado das instruções pelos maquinistas.

Tabela 2 - User Story 2 - Monitorar Estações e Manutenção

Número	2
Título	Monitorar Estações e Manutenção
Descrição	<p>Como gestor de operações, eu quero monitorar as estações e a manutenção programada, para garantir que tudo esteja funcionando conforme o planejado.</p>
Critérios de Aceitação	<ol style="list-style-type: none"> 1. O status das estações deve ser atualizado automaticamente a cada 10 segundos, refletindo os estados "Operacional", "Inativo" e "Em Manutenção", com cores distintas para cada estado. 2. As informações sobre manutenção programada devem incluir a data, horário, estação afetada e impacto estimado.

Número	2
Testes de Aceitação	<p>Critério de Aceitação 01: O gestor visualiza o status atualizado das estações.</p> <ul style="list-style-type: none"> - Atualização visível e precisa = correto. - Atualização ausente ou atrasada = incorreto, precisa de ajuste. <p>Critério de Aceitação 02: O gestor vê informações completas sobre manutenções programadas.</p> <ul style="list-style-type: none"> - Informações completas exibidas = correto. - Informações incompletas ou ausentes = incorreto, precisa de ajuste.
Prioridade	Alta (Must Have). Monitorar o status das estações em tempo real é essencial para a operação contínua e para evitar falhas críticas que impactem o serviço.

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

A User Story acima atende ao modelo INVEST. Ela é **independente**, pois permite o monitoramento das estações sem depender de outras funcionalidades. É **negociável**, com flexibilidade para ajustar o nível de detalhamento do monitoramento em tempo real. A User Story é **valiosa** ao proporcionar ao gestor visibilidade contínua, ajudando a identificar falhas ou atrasos que possam comprometer a operação. Além disso, é **estimável**, pois seu escopo é claro e facilita a avaliação de esforço. A história é **pequena**, abordando exclusivamente o monitoramento das estações, e **testável**, com critérios claros para verificar se as informações de status e manutenção estão visíveis em tempo real.

Tabela 3 - User Story 3 - Alocar Trens e Maquinistas

Número	3
Título	Alocar Trens e Maquinistas
Descrição	<p>Como gestor de operações, eu quero alocar trens e maquinistas de forma eficiente, para garantir que o serviço seja mantido sem interrupções.</p>
Critérios de Aceitação	<ol style="list-style-type: none"> 1. O sistema deve exibir a disponibilidade de trens e maquinistas em tempo real, incluindo dados como horários livres e restrições. 2. Em caso de indisponibilidade de recursos, o sistema deve enviar notificações automáticas ao gestor, incluindo uma sugestão de alocação alternativa.
Testes de Aceitação	<p>Critério de Aceitação 01: O gestor vê a disponibilidade em tempo real.</p> <ul style="list-style-type: none"> - Disponibilidade exibida corretamente = correto. - Disponibilidade ausente ou incorreta = incorreto, precisa de ajuste. <p>Critério de Aceitação 02: O gestor recebe notificações de indisponibilidade.</p> <ul style="list-style-type: none"> - Notificação recebida com sugestão de alternativa = correto. - Notificação ausente ou incompleta = incorreto, precisa de ajuste.
Prioridade	Alta (Must Have). Garantir a alocação eficiente de recursos é essencial para evitar interrupções no serviço e melhorar a experiência do usuário.

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

A User Story 3 também segue o modelo INVEST. Ela é **independente**, pois aborda exclusivamente a alocação de trens e maquinistas, sem exigir outras funcionalidades. É **negociável**, permitindo ajustes na apresentação dos dados de disponibilidade conforme as necessidades do gestor. A história é **valiosa** para o usuário, garantindo que os recursos estejam alocados de forma eficiente, evitando interrupções no serviço. Além disso, é **estimável**, com escopo bem delimitado, facilitando a avaliação do esforço necessário para implementá-la. A User Story é **pequena**, focando apenas na visibilidade e alocação de trens e maquinistas, e **testável**, com critérios verificáveis que permitem confirmar a visibilidade em tempo real e a funcionalidade de alocação conforme a necessidade operacional.

Tabela 4 - User Story 4 - Verificar Regularidade do Plano PPR

Número	4
Título	Verificar Regularidade do Plano PPR
Descrição	Como gestor de operações, eu quero verificar a regularidade do Plano PPR, para assegurar que ele esteja sendo cumprido adequadamente.
Critérios de Aceitação	1. O sistema deve gerar relatórios automáticos semanais e mensais, destacando métricas como porcentagem de cumprimento e atrasos identificados. 2. O relatório deve estar disponível em formato PDF e CSV, com opções de filtros por período e setor.
Testes de Aceitação	Critério de Aceitação 01: Relatório gerado com métricas detalhadas. - Relatório disponível e preciso = correto. - Relatório ausente ou impreciso = incorreto, precisa de ajuste.
Prioridade	Média (Should Have). Monitorar o cumprimento do Plano PPR é importante para manter a conformidade com os objetivos operacionais e evitar atrasos nos processos.

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

A User Story 4 acima segue o modelo INVEST. Ela é **independente**, pois trata exclusivamente da verificação do cumprimento do Plano PPR sem necessitar de outras funcionalidades. A história é **negociável**, com possibilidade de ajustar a frequência e o nível de detalhe dos relatórios gerados. É **valiosa**, pois garante que o gestor possa monitorar e assegurar o cumprimento do plano, evitando desvios operacionais. Além disso, a User Story é **estimável**, pois tem escopo bem definido, facilitando a estimativa de esforço. A história é **pequena**, focando apenas no cumprimento do plano, e **testável**, com critérios claros para verificar a geração e a precisão dos relatórios automáticos sobre o plano.

Tabela 5 - User Story 5 - Usabilidade fácil de Dashboard

Número	5
Título	Usabilidade fácil de Dashboard
Descrição	Como gestor de operações, eu quero usar o Dashboard com muita facilidade, para tirar o máximo proveito das informações dele e conseguir me comunicar com meu time de forma eficaz.

Número	5
Critérios de Aceitação	<p>1. O Dashboard deve ter uma interface intuitiva, com botões, filtros e gráficos organizados de forma clara e acessível.</p> <p>2. O tempo de carregamento de qualquer funcionalidade não deve exceder 3 segundos.</p>
Testes de Aceitação	<p>Critério de Aceitação 01: O gestor consegue acessar e utilizar os filtros e gráficos intuitivamente.</p> <ul style="list-style-type: none"> - Funcionalidades acessíveis e fáceis de usar = correto. - Dificuldade de navegação = incorreto, precisa de ajuste. <p>Critério de Aceitação 02: O tempo de carregamento é inferior a 3 segundos.</p> <ul style="list-style-type: none"> - Tempo dentro do limite = correto. - Tempo excede o limite = incorreto, precisa de ajuste.
Prioridade	Alta (Must Have). A usabilidade do Dashboard é essencial para garantir que as informações sejam aproveitadas de forma eficiente e contribuam para decisões rápidas e precisas.

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

A User Story 5 atende ao modelo *INVEST*. Ela é **independente**, pois permite que o gestor gerencie cautelas e instruções sem depender de outras funcionalidades do sistema. A história é **negociável**, já que o formato e a frequência de envio das instruções podem ser ajustados conforme as necessidades operacionais. É **valiosa**, pois assegura que os maquinistas recebam informações críticas de segurança, contribuindo para operações mais seguras e organizadas. A User Story é **estimável**, dado que seu escopo claro facilita a avaliação do esforço necessário para implementá-la. Ela é **pequena**, abrangendo exclusivamente o envio e o gerenciamento de cautelas e instruções, e **testável**, com critérios que permitem verificar se as instruções foram devidamente enviadas e recebidas pelos maquinistas.

Tabela 6 - User Story 6 - Registrar Ocorrências e Falhas

Número	6
Título	Registrar Ocorrências e Falhas
Descrição	Como gestor de operações, eu quero registrar ocorrências e falhas, para que eu possa acompanhar e resolver os problemas de forma eficaz.
Critérios de Aceitação	<p>1. O sistema deve permitir o registro detalhado de ocorrências, incluindo dados como tipo de falha, descrição, horário e responsável pela resolução.</p> <p>2. Relatórios automáticos semanais devem incluir análise de ocorrências recorrentes e métricas de resolução.</p>

Número	6
Testes de Aceitação	<p>Critério de Aceitação 01: Ocorrências registradas com todos os campos obrigatórios preenchidos.</p> <ul style="list-style-type: none">- Registro completo e detalhado = correto.- Registro incompleto ou ausente = incorreto, precisa de ajuste.
Prioridade	<p>Alta (Must Have). Monitorar e resolver falhas de maneira eficaz é fundamental para garantir a operação contínua e segura do sistema.</p>

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Por fim, a User Story acima também não falha em seguir o modelo *INVEST*. Ela é **independente**, tratando exclusivamente do registro de ocorrências e falhas, sem depender de outras funcionalidades para implementação. A história é **negociável**, permitindo ajustes na forma de categorização e no conteúdo dos relatórios periódicos. É **valiosa**, pois proporciona ao gestor um recurso essencial para monitoramento contínuo, facilitando a identificação e resolução de problemas operacionais. A User Story é **estimável**, pois é específica e tem um escopo bem delimitado, permitindo uma estimativa de esforço. Ela é **pequena**, focada apenas no registro e acompanhamento de falhas, e **testável**, com critérios claros para verificar a precisão e a frequência dos registros e relatórios gerados.

A construção de User Stories é fundamental para garantir que a solução desenvolvida atenda às necessidades reais dos usuários. A partir das personas criadas e das interações mapeadas, foi possível elaborar histórias que guiam o desenvolvimento de funcionalidades focadas em resolver os principais desafios enfrentados pelo gestor de operações. Além disso, ao garantir que as histórias estejam conectadas a critérios de aceitação claros, o projeto proporciona uma base sólida para o sucesso da solução, assegurando que ela atenda às expectativas e objetivos do cliente final.

1.4. Wireframe

Um wireframe é um diagrama visual que esboça a estrutura de uma tela de um site ou aplicativo, demonstrando como os elementos se relacionam entre si e como são estruturados. Devido a seu caráter inicial nos estágios de um projeto, o wireframe pode ser pensado como um rascunho para validar uma ideia e, a partir desta validação, criar a primeira versão de um protótipo.

Em nosso projeto com a CPTM temos como objetivo criar um infográfico para visualização de dados que apoiem a tomada de decisões da operação, assim, urge a necessidade de um wireframe. ([Miro, 2024](#))

Para garantir uma visualização clara e eficaz dos dados operacionais da CPTM, selecionamos gráficos específicos que transmitem informações essenciais de cada situação. A escolha de cada gráfico foi fundamentada na necessidade de facilitar a leitura e a interpretação dos dados pelo usuário final, considerando o perfil de cada público-alvo.

Utilizamos o gráfico de bolhas para destacar as interrupções e falhas de sensores por linha. Esse tipo de gráfico permite uma visualização rápida do volume de falhas, onde o tamanho das bolhas representa a quantidade de falhas registradas, e as cores diferenciam as linhas atendidas pelos trens. Essa representação visual oferece uma identificação imediata dos setores que exigem atenção, facilitando a tomada de decisões pela operação geral e diretoria.

Para identificar o trem com o maior número de falhas, optamos pelo gráfico de barras verticais. Nesse gráfico, o eixo X representa cada trem e o eixo Y indica a quantidade de falhas. Essa forma de visualização é ideal para a operação local, pois permite uma comparação direta e clara entre os trens, ajudando a priorizar manutenções e verificar padrões de falhas em trens específicos. Além disso, detalhamos o motivo de priorização deste tipo de gráfico: a comparação direta facilita decisões rápidas, reforçando o porquê de termos escolhido barras verticais em vez de outras representações.

Para analisar os dias da semana e horários com maior pico de pessoas por estação, escolhemos o heatmap (mapa de calor). Nele, os dias da semana estão no eixo X, os horários no eixo Y, e as cores representam o número de passageiros. Esse gráfico facilita a identificação visual dos picos de movimentação, permitindo que a operação local ajuste recursos e pessoal de acordo com os períodos de maior demanda. Destacamos também que essa escolha se deve à necessidade de evidenciar padrões temporais de modo simples, justificando a priorização do heatmap como ferramenta de análise do fluxo de passageiros.

Ao observar as datas anuais com maior intensidade de passageiros de maneira geral, utilizamos o gráfico de dispersão. Nesse caso, o eixo X representa o dia, o mês e o ano, o eixo Y mostra o fluxo de passageiros, e cada ponto corresponde a datas específicas. Essa visualização é útil para a diretoria, pois evidencia tendências sazonais e auxilia no planejamento estratégico para períodos de alta demanda.

Para destacar as estações com maior fluxo de passageiros, novamente empregamos o gráfico de barras verticais, onde o eixo X representa as estações e o eixo Y indica o fluxo diário. Essa escolha facilita a visualização direta das estações mais movimentadas, auxiliando a diretoria no planejamento de melhorias de infraestrutura e alocação de recursos.

Na análise da relação entre o tempo de abertura e fechamento de portas e o volume de passageiros em diferentes horários, utilizamos outro gráfico de bolhas. Nesse gráfico, o eixo X mostra o volume de passageiros, o eixo Y representa o tempo de abertura/fechamento das portas, e o tamanho das bolhas indica o horário (como picos ou não). Essa visualização ajuda a operação local a entender como o fluxo de passageiros afeta o tempo de operação das portas, permitindo otimizações operacionais.

Por fim, para comparar as linhas em termos de falhas e demanda simultaneamente, selecionamos um gráfico empilhado. Nele, as falhas e a demanda aparecem empilhadas, mostrando o total da carga em cada linha. Essa representação é valiosa para a diretoria, pois permite uma compreensão integrada dos desafios operacionais de cada linha, orientando decisões sobre investimentos e priorização de recursos, além de ajudar a diretoria a correlacionar falhas com picos de demanda.

Para aumentar a eficácia e a interação dos dashboards e wireframes, aplicamos algumas técnicas avançadas. A antecipação de ação do usuário foi incorporada para prever necessidades e exibir dados críticos proativamente. Por exemplo, o sistema pode gerar alertas automáticos sobre falhas inesperadas ou variações significativas no fluxo de passageiros, permitindo que o usuário tome decisões de forma proativa e reduza o tempo de resposta a incidentes. Microinterações, como respostas visuais e animações sutis, também foram aplicadas para enriquecer a experiência do usuário. Por exemplo, ao passar o cursor sobre gráficos ou tocar em elementos, detalhes adicionais aparecem dinamicamente, incentivando a interação sem sobrecarregar o

design. Affordances também foram integradas, com elementos visuais claros que indicam ao usuário como interagir com o sistema, como botões destacados ou ícones intuitivos.

Para ilustrar a estrutura e a aplicação desses gráficos, desenvolvemos wireframes em baixa fidelidade voltados para dashboards e dispositivos móveis. Esses wireframes representam a disposição e os elementos de design planejados, servindo como base para o protótipo final. Eles consideram tanto a visualização em dispositivos desktop quanto móveis, garantindo que o design seja responsivo e otimizado para telas menores, com consideração para toques e gestos típicos de dispositivos móveis. Ao receber opiniões externas de usuários-teste, revisamos os wireframes, ajustando a posição de certos gráficos e acrescentando marcadores visuais que indicam a possibilidade de interação, detalhando assim os ajustes específicos feitos após as opiniões recebidas.

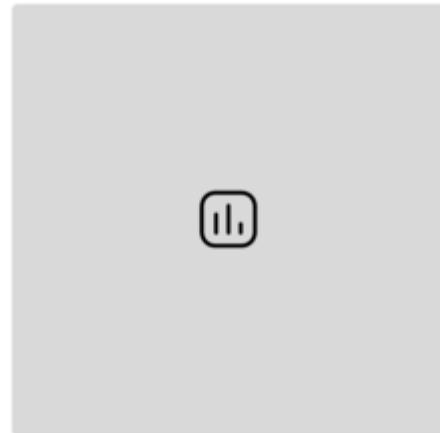
Os wireframes têm como principal objetivo demonstrar a integração e organização dos gráficos. Ou seja, cada gráfico foi posicionado estratégicamente para facilitar o fluxo de leitura dos dados, priorizando informações críticas e garantindo que o usuário possa navegar facilmente entre diferentes seções do dashboard.

As imagens dos wireframes estão incluídas a seguir, ilustrando como cada gráfico será apresentado e como as técnicas avançadas foram incorporadas ao design. Essas representações visuais auxiliam na comunicação do fluxo e da interface da solução proposta, assegurando que o design atenda aos objetivos do usuário e ofereça uma experiência intuitiva e eficiente. É importante ressaltar que, nos wireframes, não estão incluídos exatamente os gráficos que serão utilizados, mas sim uma ideia de onde eles estarão inseridos, já que esse é um wireframe de baixa fidelidade.

Figura 6 - Wireframe Desktop

Título 1

Nome do Gráfico 1

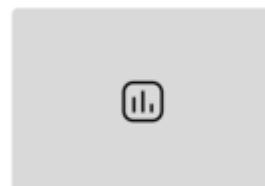


Placeholder text for the first critical section.

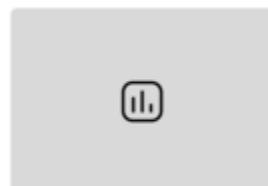
Placeholder text for the first critical section.

Placeholder text for the first critical section.

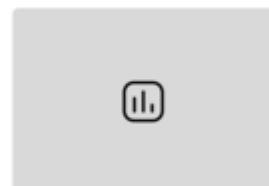
Nome do Gráfico 2



Nome do Gráfico 3



Nome do Gráfico 4

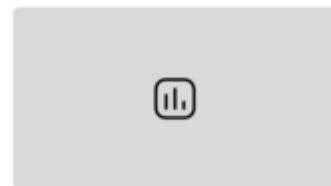


Título 2

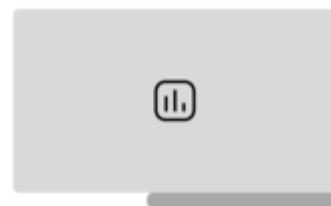
Nome do Gráfico 5



Nome do Gráfico 7



Nome do Gráfico 8



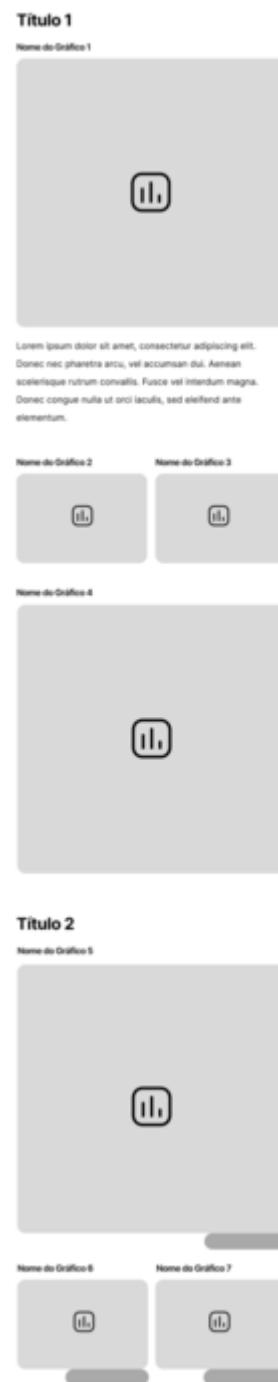
Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Primeiramente, temos o wireframe adaptado para a versão desktop, provavelmente a que mais será utilizada pelos funcionários da CPTM. Ele é dividido em duas seções principais, separadas pelos "Título 1" e "Título 2". A primeira parte é mais voltada para as áreas de operação geral e diretoria da empresa, contendo informações mais críticas, como interrupções e falhas de sensores, destacados em um gráfico de bolhas grande. Os gráficos menores também contém informações importantes, porém não tão críticas.

Já a segunda parte é voltada para a operação local, contendo informações relacionadas ao fluxo de passageiros, por exemplo, inseridas em um mapa de calor de tamanho maior. Os outros 2 menores gráficos, assim como na seção anterior, também são importantes, mas não necessitam de uma atenção imediata. Além disso, essa segunda seção contém uma barra cinza escura embaixo dos gráficos. Isso serve para que o usuário

da plataforma consiga ver as informações para cada linha de trem diferente, por exemplo analisando se a Linha Safira ou a Linha Esmeralda têm maior fluxo de passageiros por dia, semana, mês ou ano. Como resultado de feedbacks externos, acrescentamos indicadores visuais e instruções sobre como interagir com essa barra, tornando claro ao usuário como alternar entre linhas e períodos, detalhando assim as melhorias feitas com base em opiniões externas.

Figura 7 - Wireframe Mobile



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Além disso, foi desenvolvido um wireframe adaptado para dispositivos móveis, denominado "wireframe mobile". Essa versão mantém a estrutura e os elementos da versão desktop, porém otimizada para o formato e usabilidade em telas menores, garantindo uma experiência fluida e intuitiva no mobile. Além disso, um dos gráficos da primeira seção tem o tamanho maior, por ser ligeiramente mais importante que os outros menores, e também para que o layout da página fique melhor.

Ao integrar cuidadosamente a escolha dos gráficos com técnicas avançadas de design e usabilidade, o dashboard proposto oferecerá uma ferramenta poderosa para a gestão operacional da CPTM, atendendo às necessidades específicas dos diferentes públicos-alvo e facilitando a tomada de decisões estratégicas.

2. Conclusões

Até aqui, o que começou como um esforço para entender o usuário e suas necessidades já se transformou em um plano de ação estruturado. Nós não apenas definimos uma persona – dando um rosto, contexto e motivações à figura do gestor operacional Sérgio – mas também mapeamos a jornada real desse usuário, mostrando o caminho que ele percorre, as dificuldades que enfrenta e onde podemos intervir. A partir disso, criamos user stories focadas no que realmente importa para a operação, garantindo que cada funcionalidade desenvolvida atenda a um propósito claro.

Além disso, os wireframes serviram como um primeiro esboço visual, permitindo visualizar a integração dos dados e funcionalidades num formato simples e intuitivo. Essas representações ajudaram a antecipar problemas, guiar decisões de design e alinhar expectativas antes mesmo da fase de desenvolvimento.

Em outras palavras, o projeto não está apenas no papel: já existem bases sólidas para criar uma solução centrada no usuário, capaz de tornar dados complexos em insights práticos, melhorar rotinas e facilitar a tomada de decisão dentro da CPTM. O próximo passo é transformar esses planos em um protótipo funcional, testar, ajustar e, assim, avançar em direção a uma ferramenta que realmente faça a diferença na rotina do Sérgio e de toda a equipe operacional.

3. Referências

ARRUDA, Ricardo. O que são User Stories (Estórias de Usuário)? - Agile Expert. 14 maio 2021. Disponível em: <https://www.agileexpert.com.br/2021/05/14/o-que-sao-user-stories-historias-de-usuario/>. Acesso em: 17 out. 2024.

BABICH, N. (2017, 29 de setembro). Putting Personas to Work in UX Design: What They Are and Why They're Important. Welcome to the Adobe Blog. <https://blog.adobe.com/en/publish/2017/09/29/putting-personas-to-work-in-ux-design-what-they-are-and-why-theyre-important>. Acesso em: 17 out. 2024.

MIRO. O que é wireframe? Disponível em: <https://miro.com/pt/wireframe/o-que-e-wireframe/>. Acesso em: 10 out. 2024.

The Interaction Design Foundation. What is User Centered Design (UCD)? (2016, 5 de junho). <https://www.interaction-design.org/literature/topics/user-centered-design>. Acesso em: 17 out. 2024.

Documentação da parte de Programação do Projeto Big Data - Módulo 8 - Inteli

Grupo Pérola Negra - Solução DataApp com Dashbord

Integrantes do Grupo:

- Ana Martire
- Eduardo Oliveira
- Keylla Oliveira
- Lucas Barbosa
- Nicollas Isaac
- Sophia Nóbrega

Sumário

- Documentação da parte de Programação do Projeto Big Data - Módulo 8 - Inteli
 - Grupo Pérola Negra - Solução DataApp com Dashbord
 - Integrantes do Grupo:
- Sumário
- 1. Data Product Canvas
 - Problema
 - Dados
 - Solução
 - Hipóteses
 - KPIs
 - Atores
 - Ações
 - Valores
 - Riscos
 - Performance/Impacto
- 2. Arquitetura Macro
 - 2.1. Componentes da Arquitetura
 - 2.2 UML de Componentes
 - Arquitetura Medallion
 - Camada de Bronze (Data Lake)
 - Camada de Prata (Transformação e Normalização)
 - Camada de Ouro (Data Warehouse)
 - Camada de Ródio (Visualização)
 - Ingestão
 - Transformação
 - Análise
 - Visualização
- 3. Processo de ETL

- ETL
 - Arquitetura e Fluxo do Pipeline de Dados
 - Resumo do Pipeline
 - 3.1 Análise de Dados
 - 3.1.1 Criação das Planilhas
 - 3.2 Cubo de Dados
- 4. Análise de Impacto Ético
 - Introdução
 - Impactos em Meio Ambiente e Sociedade
 - 4.1. Privacidade e Proteção de Dados
 - 4.2. Equidade e Justiça
 - 4.3. Transparência e Consentimento Informado
 - 4.4. Responsabilidade Social
 - 4.5. Viés e Discriminação
 - 4.6. Responsabilidade social
 - Conclusão
- 5. Streamlit e Infográfico
 - 5.1. Documentação do Streamlit
 - 5.1.1. Autenticação
 - 5.1.2. Dashboard
 - 5.2. Documentação dos Filtros
 - 5.3. Documentação do Infográfico
 - 5.4 Documentação dos Relatórios
- 6. Cobertura de Testes
 - 6.1. Objetivo
 - 6.2 Estrutura de Testes
 - 1. Testes para Processos ETL
 - Casos Testados
 - Ferramentas Utilizadas
 - 2. Testes para Views
 - Casos Testados
 - Integração com Streamlit
 - 3. Geração de Relatórios
 - Passos para Geração do Relatório
 - 6.3. Conexões no Streamlit para Testes
 - 6.4. Conclusão
- 7. Conclusões e Próximos Passos
 - 7.1. Conclusões Obtidas
 - 7.2. Próximos Passos
 - 7.3. Outras Perspectivas e Ideias Futuras
- 8. Anexos
 - Anexo I
- Termo de Uso e Política de Privacidade de Dados
 - 1. Disposições Gerais
 - 2. Objetivo da Coleta de Dados
 - 3. Tipos de Dados Coletados

- 4. Consentimento e Revogação
 - 4.1 Obtenção de Consentimento
 - 4.2 Revogação de Consentimento
- 5. Transparência e Acesso à Informação
- 6. Segurança e Armazenamento de Dados
- 7. Direitos dos Usuários
- 8. Auditoria e Conformidade
- 9. Canal de Atendimento ao Usuário
- 10. Disposições Finais
- 9. Automatização de Coleta
 - 9.1. Automatização do Data Ingestion
 - 9.2. Conclusão
- 10. Referências

1. Data Product Canvas

Dividido em 10 blocos (problema, solução, dados, hipóteses, atores, ações, KPIs, valores, riscos e performance/impacto), o Data Product Canvas é um framework desenvolvido para auxiliar no planejamento estratégico e execução de produtos de dados. Baseado na Metodologia Ágil/Lean, ele organiza as informações essenciais do projeto em um modelo visual, alinhando a visão de todos os stakeholders. Seu principal objetivo é proporcionar clareza sobre o propósito do projeto e facilitar a geração de um roadmap estruturado.

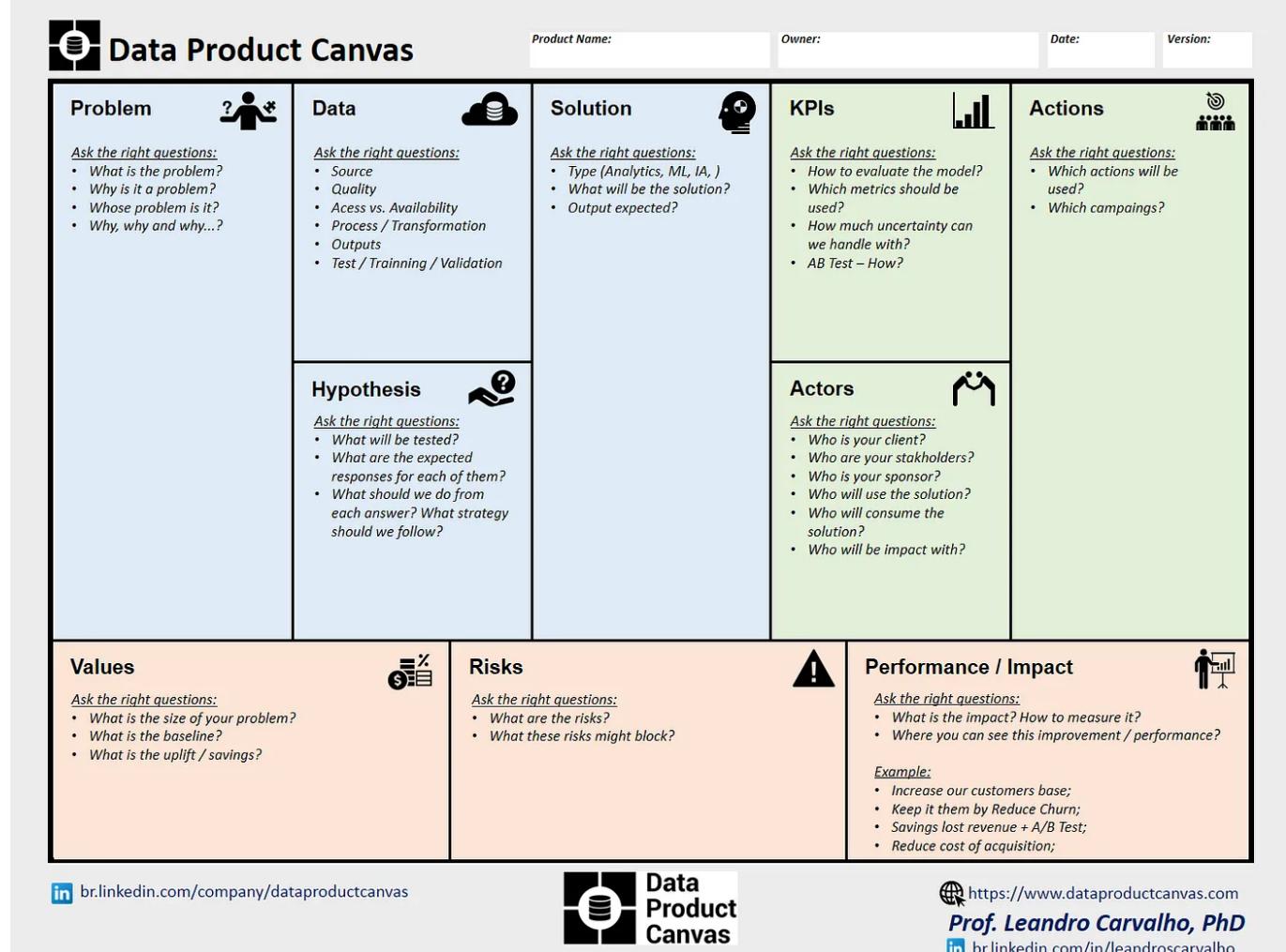
Cada bloco do Canvas é distribuído em 3 áreas de domínio:

1. **Visão do produto:** composta pelos blocos problema, solução, dados e hipóteses;
2. **Visão da estratégia:** composta pelos blocos atores, ações e KPIs;
3. **Visão do negócio:** composta pelos blocos valores, riscos e performance/impacto.

O uso do Data Product Canvas no projeto da CPTM destaca sua aplicação prática na resolução de problemas operacionais críticos. O framework foi utilizado para mapear as necessidades de eficiência e segurança no transporte público, priorizando o monitoramento automatizado de operações ferroviárias, como falhas captadas pelos sensores, fluxo de passageiros e sincronização de portas.

Além disso, a proposta de valor do produto está diretamente alinhada às demandas específicas da CPTM. O pipeline desenvolvido permite ações rápidas e corretivas, reduzindo o impacto das falhas e melhorando o atendimento ao cliente. Por exemplo, com notificações em tempo real, operadores e técnicos conseguem identificar e corrigir problemas antes que afetem o fluxo de passageiros, garantindo maior confiabilidade na operação.

A imagem abaixo apresenta o template do Data Product Canvas, demonstrando sua estrutura visual e organização em blocos. Cada bloco é preenchido com informações que detalham o Discovery necessário para uma compreensão única de cada parte do produto de dados que será desenvolvido.

Figura 1 - Template Data Product Canvas

Fonte: Leandro Carvalho (2024)

Em cada um é explorado em detalhes todo o Discovery necessário para que se tenha um entendimento único de cada parte do produto de dados que será desenvolvido. E cada domínio trata de uma área chave para o correto planejamento e desenvolvimento do produto, fornecendo uma visão 360 que vai da determinação do problema até a execução estratégica, passando pelo monitoramento de KPIs e mapeamento dos riscos. (Carvalho, 2024)

Figura 2 - DPC Caixa Preta



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Problema

O desafio principal está relacionado à incapacidade de detectar e reagir rapidamente a falhas ou anomalias operacionais captadas pela caixa preta dos trens. A análise manual dos dados pode ser lenta e propensa a erros, levando a ineficiências operacionais, falhas não identificadas e potenciais riscos à segurança.

Dados

O dataset inclui informações de diversas tabelas, principalmente relacionadas aos sensores dos trens e às operações da CPTM. Os mesmos são importantes para monitorar o comportamento do trem e identificar possíveis anomalias operacionais, como falhas de sensor, uso inadequado de portas e fluxo irregular de passageiros. Abaixo está o detalhamento das colunas de cada tabela:

Figura 3 - Schema SQL Tabelas

sua_tabela	dmo_anl_vw_tot_mov_periodo	dmo_anl_vw_tipo_embarque	dmo_anl_vw_intervalos_dia	users
int No datetime Open_Time datetime Closed_Time int Line_ID int Train_ID int StartStation_ID int Station_ID int NextStation_ID int EndStation_ID int Carriage_ID int Door_ID int IN int OUT int Command int SensorSts string filename float Door_Open_Duration int Hour category Time_Interval int Day_of_Week	int id_dt_hora_minuto int cod_bilh int cd_estac_bu datetime dt_validacao int total_validades category tipo_dia	int id_tipo_embarque category tx_movimento int cod_bilh int id_tipo_lancamento_fk category tx_lancamento	datetime dt_hora_minuto int id_dt_hora_minuto string hora_ini string hora_fim category tx_prefixo	category id category name category email float passwordhash int id_estacao int id_stacao category tx_nome int id_estacao_bu

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Abaixo encontra-se uma descrição dos dados de cada tabela fornecida pela CPTM. Vale ressaltar que tais interpretações foram realizadas pelos membros do grupo e não correspondem a uma descrição oficial dos dados.

- **Tabela: sua_tabela**

- **No** (int): Identificador único do registro.
- **Open_Time** (datetime): Horário de abertura da porta.
- **Closed_Time** (datetime): Horário de fechamento da porta.
- **Line_ID** (int): Identificador da linha do trem.
- **Train_ID** (int): Identificador do trem.
- **StartStation_ID** (int): Identificador da estação inicial.
- **Station_ID** (int): Identificador da estação atual.
- **NextStation_ID** (int): Identificador da próxima estação.
- **EndStation_ID** (int): Identificador da estação final.
- **Carriage_ID** (int): Identificador do vagão.
- **Door_ID** (int): Identificador da porta.
- **IN** (int): Quantidade de passageiros que entraram.
- **OUT** (int): Quantidade de passageiros que saíram.
- **Command** (int): Comando acionado para abertura ou fechamento.
- **SensorSts** (int): Status do sensor.
- **filename** (string): Nome do arquivo do log de dados.
- **Door_Open_Duration** (float): Duração da abertura da porta.
- **Hour** (int): Hora do evento.
- **Time_Interval** (category): Intervalo de tempo categorizado.
- **Day_of_Week** (int): Dia da semana.

- **Tabela: dmo_anl_vw_mov_periodo**

- **id_dt_hora_minuto** (int): Identificador da data e hora.
- **cod_bilh** (int): Código do bilhete.
- **cd_estac_bu** (int): Código da estação.
- **dt_validacao** (datetime): Data de validação.
- **total_validacoes** (int): Total de validações de bilhetes.
- **categoria** (category): Categoria de bilhete.

- **Tabela: dmo_anl_vw_tipo_embarque**

- **cd_tipo_embarque** (int): Código do tipo de embarque.
- **tx_movimento** (category): Tipo de movimento (entrada/saída).
- **cod_bilh** (int): Código do bilhete.
- **cd_tipo_lancamento_fk** (int): Código de lançamento.
- **tx_lancamento** (category): Tipo de lançamento.

- **Tabela: dmo_anl_vw_intervalos_dia**

- **dt_hora_minuto** (datetime): Data e hora.
- **id_dt_hora_minuto** (int): Identificador do minuto.
- **hora_ini** (string): Hora inicial do intervalo.

- **hora_fim** (string): Hora final do intervalo.
- **Tabela: dmo_anl_vw_estacoes**
 - **id_estacao** (int): Identificador da estação.
 - **tx_prefixo** (category): Prefixo da estação.
 - **tx_nome** (category): Nome da estação.
 - **cd_estacao_bu** (int): Código da estação.

- **Tabela: users**

- **id** (category): Identificador do usuário.
- **name** (category): Nome do usuário.
- **email** (category): E-mail do usuário.
- **passwordhash** (float): Hash da senha para autenticação.

Solução

Um pipeline automatizado para análise dos dados, no caso do grupo Pérola Negra, focado na caixa preta, com o objetivo de detectar e monitorar anomalias, além de otimizar as operações de embarque e desembarque. A solução pode envolver:

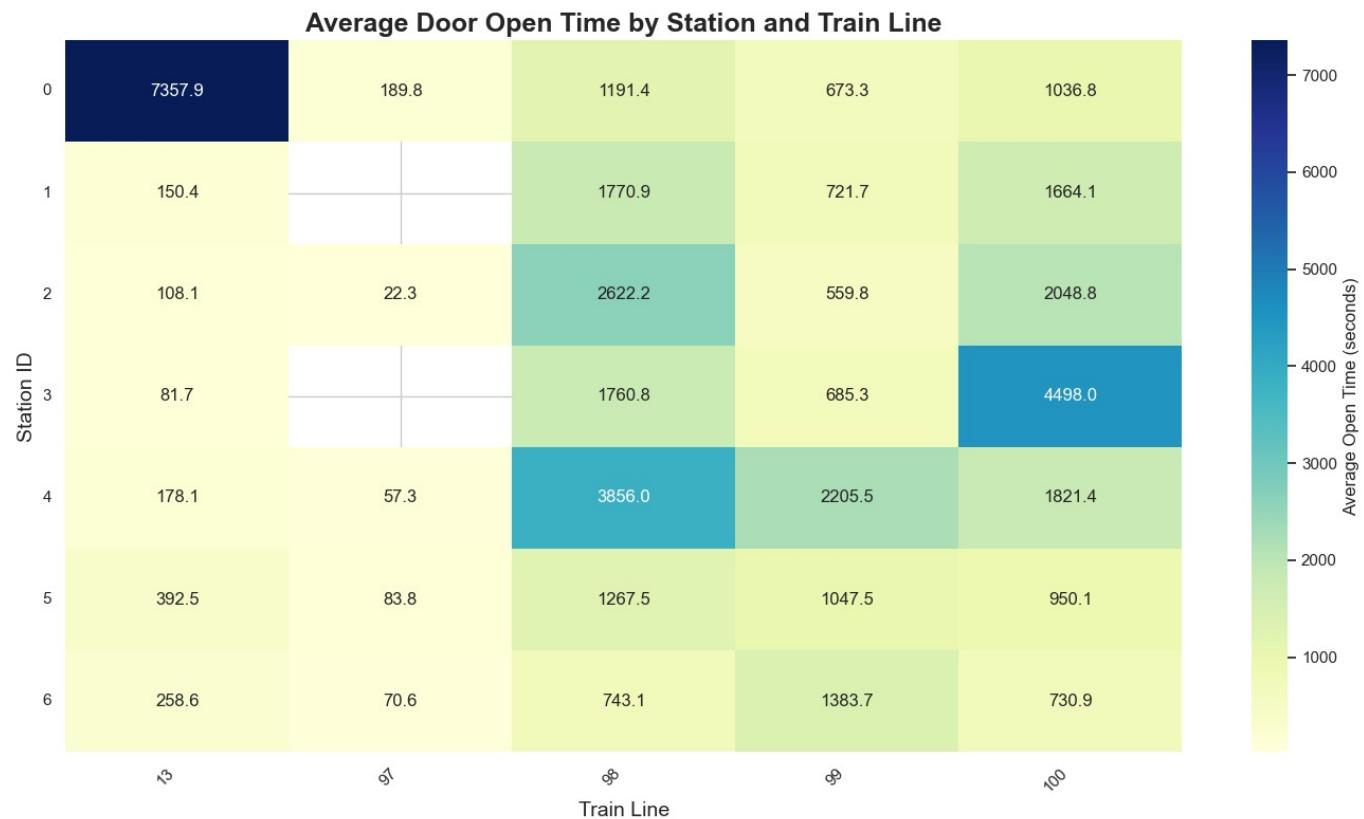
- Monitoramento contínuo de falhas nos sensores e comandos;
- Identificação de padrões operacionais críticos;
- Otimização da eficiência do fluxo de passageiros e alocação de frota com base nos dados históricos.

Hipóteses

- **H1:** A movimentação de passageiros (**IN** e **OUT**) varia significativamente com o horário do dia e a estação, sendo maior nas horas de pico e em estações centrais.
- **H2:** Certas portas são mais utilizadas, especialmente em vagões ou posições específicas do trem, o que pode impactar a eficiência operacional do embarque/desembarque.
- **H3:** Anomalias na sincronização das portas (diferença entre abertura e fechamento) impactam diretamente o tempo de parada nas estações.

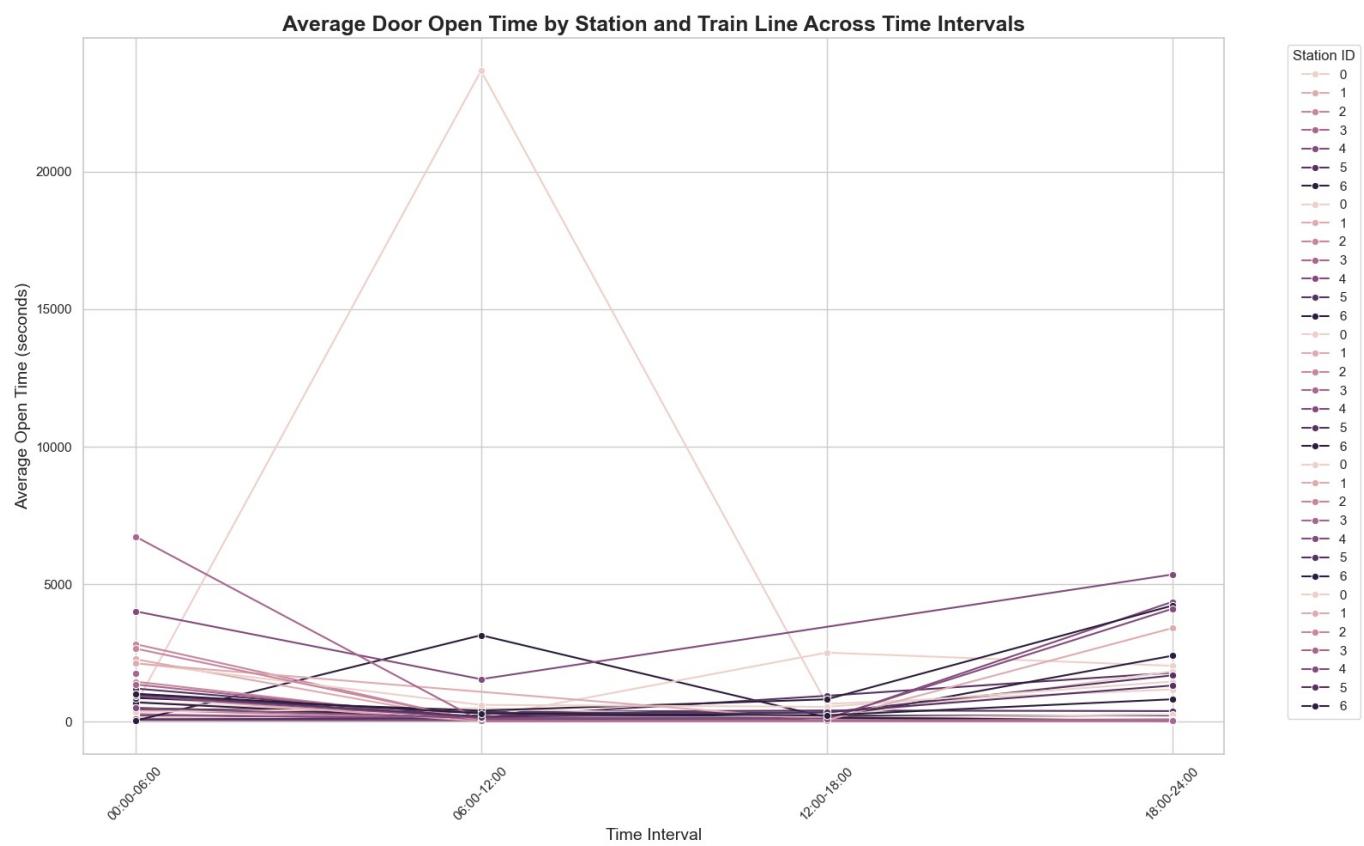
As hipóteses foram formuladas com base na análise exploratória realizada pela equipe. Identificamos padrões significativos através dos gráficos apresentados a seguir, que ilustram as correlações mencionadas:

Gráfico 4 - Média Abertura da Porta por Estação e Linha



Fonte: Leandro Carvalho (2024)

Gráfico 5 - Média Abertura da Porta por Estação e Linha com Intervalos de Tempo



Fonte: Leandro Carvalho (2024)

KPIs

1. **Monitoramento de embarque/desembarque por porta:** Avaliar a eficiência das portas durante o embarque e desembarque em diferentes horários e estações. Isso envolve medir o fluxo de passageiros (**IN** e **OUT**) por porta (**Door_ID**) e identificar padrões de uso. Este KPI ajuda a entender se certas portas ou vagões são subutilizados, o que pode guiar otimizações no layout dos trens e das plataformas.
2. **Controle da frota de carros por horários de pico:** Monitorar a quantidade de trens alocados em períodos de maior demanda (horários de pico) e garantir que a frota seja dimensionada adequadamente para reduzir a superlotação. A eficiência operacional pode ser medida verificando se o número de passageiros por trem é otimizado em horários críticos, melhorando a experiência do usuário e reduzindo custos.
3. **Análise de ocorrência de falhas com regressão linear:** Aplicar regressão linear para detectar padrões e prever falhas nos sensores e comandos. Este KPI busca identificar se há variáveis que indicam maior probabilidade de falhas (como o tipo de comando enviado ao trem, **Command**, e os problemas detectados nos sensores, **SensorSts**). O objetivo é antecipar e evitar falhas futuras, melhorando a segurança e a eficiência operacional.
4. **Aplicar melhorias no fluxo de passageiros em portas menos utilizadas:** Identificar padrões de movimentação de passageiros entre portas e vagões, otimizando tanto logística quanto layout do trem para melhorar o fluxo nas áreas menos movimentadas.
5. **Redução do tempo médio de inatividade dos trens:** Monitorar e reduzir o tempo necessário para diagnosticar e corrigir falhas operacionais, garantindo que os trens voltem à operação o mais rápido possível. Este KPI mede a eficiência do sistema em agilizar processos críticos.
6. **Aumento da pontualidade operacional:** Mensurar a quantidade de viagens que são concluídas no horário previsto, refletindo melhorias na eficiência geral do sistema. Esse KPI avalia diretamente o impacto das intervenções na confiabilidade do serviço.
7. **Diminuição do índice de falhas não detectadas:** Acompanhar a redução de falhas que passam despercebidas pelo monitoramento automático, destacando o desempenho do sistema em identificar e resolver problemas antes que eles afetem a operação.
8. **Satisfação do cliente:** Avaliar a experiência dos passageiros por meio de questionários, feedback ou análise de reclamações, garantindo que as melhorias operacionais atendam às expectativas do público.

Atores

- **Persona Principal:**
 - **Nome:** Sérgio Ribeiro
 - **Idade:** 52 anos
 - **Profissão:** Gestor Operacional da CPTM, engenheiro de produção pela Escola Politécnica da USP
 - **Perfil:** Com 25 anos de experiência no setor ferroviário, Sérgio é conhecido por sua habilidade em otimizar operações e melhorar a eficiência na CPTM. Ele acredita que a análise de grandes volumes de dados é fundamental para melhorar a produtividade e a qualidade dos serviços, alinhando custos com a experiência do usuário.

- **Desafios:** Falta de ferramentas adequadas para analisar grandes volumes de dados, limitando a capacidade de identificar e prevenir falhas operacionais.
- **Objetivo:** Sérgio busca uma solução de Big Data que permita um monitoramento eficiente e preventivo das operações, ajudando a CPTM a se tornar uma referência em inovação tecnológica e qualidade de serviço.
- **Frase Motivadora:** "A eficiência operacional é fundamental, e a análise de dados pode nos ajudar a prever falhas antes que elas aconteçam."
- **Stakeholders:** Diretoria (stakeholder principal), membros do Conselho Administrativo, o Governo, a Secretaria de Transporte e a Gestão da CPTM (operação e manutenção).

Ações

1. **Análise de Relatórios Operacionais:** Sérgio pode utilizar os relatórios gerados pela solução para monitorar o desempenho das estações em tempo real. Isso permitirá que ele identifique rapidamente padrões de anomalias e tome decisões informadas sobre a operação, como ajustar horários de manutenção ou redirecionar trens.
2. **Implementação de Ações Corretivas:** Ao receber notificações de anomalias nas sincronizações das portas, Sérgio poderá acionar sua equipe para investigar a situação imediatamente. Ele poderá desenvolver um protocolo claro sobre como agir, garantindo que a resposta a problemas seja rápida e eficaz.
3. **Reuniões de Feedback e Melhoria Contínua:** Sérgio pode organizar reuniões regulares com sua equipe para discutir as descobertas a partir das análises e relatar como as ações corretivas impactaram a operação. Essas reuniões servirão para colher feedback e propor melhorias contínuas nos processos operacionais, garantindo que a solução se mantenha relevante e eficaz.
4. **Acompanhamento de Tendências de Fluxo:** Sérgio pode utilizar dados históricos de movimentação de passageiros para identificar mudanças no comportamento de uso das estações ao longo do tempo, como crescimento em horários não tradicionais ou redução em dias específicos. Com isso, ele pode propor ajustes estratégicos na operação e até reavaliar a necessidade de campanhas para aumentar o uso em horários de baixa demanda.
5. **Definição de Prioridades de Investimento:** Baseado nos relatórios sobre falhas frequentes por linha ou estação, Sérgio pode priorizar os recursos destinados à manutenção e atualização de equipamentos. Por exemplo, ele pode justificar a substituição de sensores com desempenho crítico ou sugerir melhorias nas estações com maior incidência de problemas.
6. **Criação de Indicadores Personalizados:** A partir da plataforma, Sérgio pode estabelecer novos KPIs que atendam a necessidades específicas, como o tempo médio de reparo após notificações de falhas ou a eficiência operacional de diferentes turnos. Esses indicadores podem ser incorporados nos relatórios para guiar novas metas de desempenho.

Valores

A implementação deste sistema visa gerar valor por meio de:

- **Redução de Custos Operacionais:** Diminuindo o número de falhas não detectadas ou corrigidas tarde, resultando em menos reparos caros ou interrupções de serviço.

- **Maior Segurança:** Detectando e corrigindo falhas operacionais antes que afetem a segurança dos passageiros e/ou da operação.
- **Eficiência Operacional:** Otimizando a movimentação de passageiros e a alocação de trens com base nos dados.

Riscos

- **Falhas na Coleta de Dados:** Se houver falhas ou problemas nos sensores, a qualidade dos dados pode ser comprometida.
- **Integração:** Pode haver dificuldades em integrar este sistema com os sistemas já existentes de monitoramento e operação da CPTM.
- **Dependência de Dados do Fabricante:** Como os dados da caixa preta são oriundos do fabricante do trem, pode haver limitações na flexibilidade e personalização da coleta de dados.

Performance/Impacto

- **Impacto no Passageiro:** Redução do tempo de espera durante embarques/desembarques, menor incidência de falhas operacionais, e uma operação geral mais eficiente e segura.
- **Eficiência Operacional:** Melhoria no agendamento de trens e na alocação de recursos, com base em dados mais precisos e completos sobre uso e performance dos trens.
- **Aumento de Receita:** Uma operação mais eficiente pode atrair mais passageiros, reduzir custos e aumentar a receita ao longo do tempo.

Em conclusão, o Data Product Canvas é uma ferramenta importante para projetos de Big Data, pois traz uma visão clara do produto a ser desenvolvido. Ele facilita a comunicação entre os stakeholders e a equipe, garantindo que as necessidades reais sejam atendidas e que todos estejam na mesma página à nível de compreensão do escopo, objetivo e propósito do projeto. Ao mapear ações estratégicas e focar na entrega de valor contínuo, o DPC minimiza o risco de que soluções inovadoras sejam subutilizadas, garantindo que as decisões sejam *data driven* e contribuam para a eficiência operacional contínua.

2. Arquitetura Macro

2.1. Componentes da Arquitetura

Nossa arquitetura é organizada em camadas, cada qual com uma função clara: reunir dados brutos, transformá-los, analisar informações e, finalmente, disponibilizá-las de forma simples e útil. Abaixo, estão descritos os principais componentes envolvidos e suas atribuições, mostrando o caminho completo que os dados percorrem até virarem insights.

- **Camada de Coleta (Bronze):** Aqui é onde tudo começa. Dados brutos. Muitas vezes desalinhados, incompletos ou em formatos distintos, eles chegam ao nosso repositório central (Data Lake). É o ponto

de entrada de informações vindas de diversas fontes, sejam elas bancos de dados internos, arquivos CSV, sistemas de monitoramento ou APIs.

- **Validação e Limpeza (Prata):** Após a coleta, entram em cena processos de ETL executados por ferramentas como AWS Glue ou AWS Lambda. Nesse estágio, asseguramos que as informações sejam consistentes e de qualidade. Dados duplicados são removidos, tipos de campos são padronizados, e validações com Pydantic garantem que nada fora do padrão chegue à etapa seguinte. Ao final, temos um conjunto de dados mais confiável e pronto para análises, ainda dentro do Data Lake.
- **Armazenamento e Modelagem (Ouro):** Agora que os dados foram refinados, eles são estruturados em um Data Warehouse otimizado para consultas analíticas. É nessa camada que entram tecnologias como o ClickHouse (ou outra solução OLAP), que aceleram e facilitam a realização de análises complexas, agregações e comparações históricas.
- **Processamento e Análise Distribuída:** Para extrair valor real dos dados, usamos o Apache Spark (rodando em AWS EMR) ou tecnologias equivalentes, capazes de processar grandes volumes de dados em paralelo. Isso significa analisar grandes quantidades de informações com rapidez, gerando estatísticas e indicadores que serão a base de insights valiosos.
- **Visualização e Integração (Ródio):** Por fim, todo esse trabalho de bastidor se materializa em dashboards e relatórios interativos. Ferramentas como o Streamlit entram em cena para dar ao usuário uma interface intuitiva, permitindo visualizar tendências, gargalos e oportunidades ocultas nos dados. É a "vitrine" do pipeline, onde gestores e analistas podem tomar decisões informadas e rápidas, sem precisar conhecer a fundo a infraestrutura por trás.

2.2 UML de Componentes

O diagrama de componentes UML descreve a organização e as interações entre os componentes de um sistema de software. O mesmo detalha como módulos, bibliotecas e outras partes do sistema se relacionam, destacando as dependências e interfaces possíveis para a comunicação. Para fins de UML 2.0, o termo "componente" refere-se a um módulo de classes que representa sistemas ou subsistemas independentes com capacidade de interagir com o restante do sistema.

Para isso, existe uma abordagem de desenvolvimento em torno de componentes: o desenvolvimento baseado em componentes (CBD). Nela, o diagrama de componentes identifica os diferentes componentes para que todo o sistema funcione corretamente. ([Lucidchart, 2024](#))

Em resumo, o UML de Componentes:

1. Imagina a estrutura física do sistema;
2. Presta atenção aos componentes do sistema e como eles se relacionam;
3. Enfatiza o comportamento do serviço quanto à interface.

O diagrama de componentes UML descreve a organização e as interações entre os componentes de um sistema de software. O mesmo detalha como módulos, bibliotecas e outras partes do sistema se relacionam, destacando as dependências e interfaces possíveis para a comunicação. Para fins de UML 2.0, o termo "componente" refere-se a um módulo de classes que representa sistemas ou subsistemas independentes com capacidade de interagir com o restante do sistema.

Para isso, existe uma abordagem de desenvolvimento em torno de componentes: o desenvolvimento baseado em componentes (CBD). Nela, o diagrama de componentes identifica os diferentes componentes para que todo o sistema funcione corretamente. ([Lucidchart, 2024](#))

Em resumo, o UML de Componentes:

1. Imagina a estrutura física do sistema;
2. Presta atenção aos componentes do sistema e como eles se relacionam;
3. Enfatiza o comportamento do serviço quanto à interface.

Arquitetura Medallion

A arquitetura medallion, ou medalhão, descreve uma série de camadas de dados que denotam a qualidade dos dados armazenados no Lakehouse.

Essa arquitetura garante a atomicidade, consistência, isolamento e durabilidade à medida que os dados passam por várias camadas de validações e transformações antes de serem armazenados em um layout otimizado para análise eficiente. Os termos bronze (bruto), prata (validado) e ouro (enriquecido) descrevem a qualidade dos dados em cada uma dessas camadas. ([Microsoft, 2024](#))

Figura 4 - Fluxo Geral do desenvolvimento da Solução



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

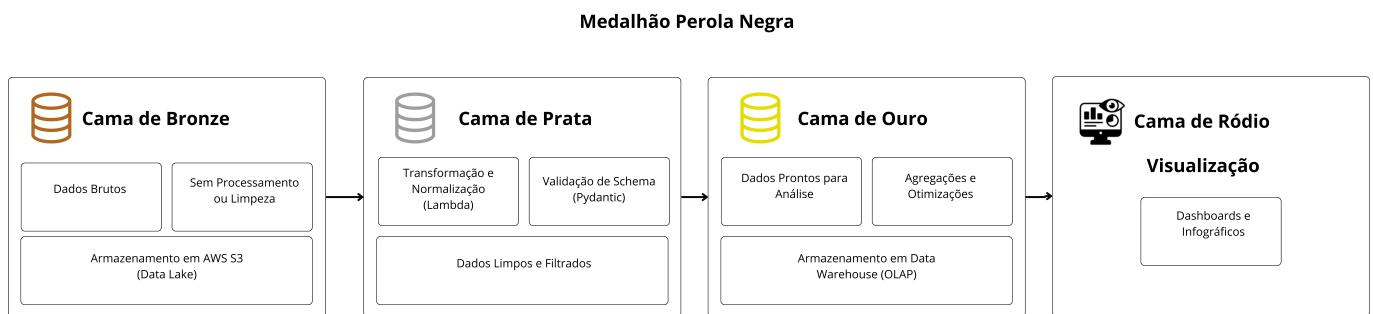
Para explicar como construímos nossa solução, acima temos uma imagem do Fluxo Geral de Desenvolvimento, sendo o mesmo composto pelas etapas:

1. **Camada de Dados**: Os dados são coletados de diversas tabelas, como Estacao, Trem_Passageiros, Intervalos_Dia e Mov_Periodo.
2. **Ingestão de Dados**: Os dados brutos obtidos das tabelas são armazenados em um repositório central.
3. **AWS S3 Data Lake**: Os dados são carregados em um data lake na AWS S3, onde passam por um processo de validação de esquema utilizando o Pydantic.
4. **Camada de Validação**: Os dados passam por processos de limpeza, deduplicação e conversão para formatos apropriados.
5. **Transformação com Lambda**: Os dados são carregados e transformados utilizando a função AWS Lambda para prepará-los para análises mais profundas.
6. **Processamento**: Os dados transformados são processados utilizando o Apache Spark, que possibilita processamento em larga escala.

7. **Data Warehouse OLAP:** Os dados processados são armazenados em um data warehouse projetado para consultas analíticas (OLAP).
 8. **Dashboards e Visualizações Streamlit:** Os resultados são apresentados por meio de dashboards interativos desenvolvidos com Streamlit.
-

Abaixo, passaremos por cada uma das camadas, bronze, prata, ouro e ródio, desenhadas para o grupo Pérola Negra considerando o foco do time nos dados de caixa preta do trem.

Figura 5 - Medalhão Perola Negra



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Camada de Bronze (Data Lake)

Armazena todos os dados de sistemas de origem externa. As estruturas da tabela desta camada correspondem às estruturas da tabela "*as is*" (como estão) no sistema de origem, juntamente com metadados de colunas adicionais, como data de carregamento, ID do processo etc.

Na Camada de Bronze do projeto "Medalhão Pérola Negra", foi trabalhado com dados fornecidos através do MinIO, uma solução de armazenamento compatível com o AWS S3 que atua como nosso Data Lake para dados brutos. Esse armazenamento inicial continha arquivos em diversos formatos, além do comum CSV. Essa diversidade de formatos exigiu um trabalho inicial de conversão e padronização.

Para garantir uma análise consistente dos dados na camada de transformação, foi necessário realizar uma etapa de conversão preliminar. Utilizando Jupyter Notebooks e ferramentas de manipulação de dados em Python, como Pandas, realizamos a extração e transformação dos dados para o formato CSV, facilitando o entendimento e permitindo uma análise exploratória mais fluida. Essa preparação inicial foi fundamental para identificar padrões, estruturar os dados e fazer as limpezas básicas necessárias antes de enviá-los para a Camada de Prata, onde são realizados os processos de transformação e validação.

Camada de Prata (Transformação e Normalização)

A Camada de Prata é onde ocorre a transformação e a limpeza dos dados. Nesta etapa, se inicia o processo de ETL (Extração, Transformação e Carga). Durante a transformação, os dados passam por normalizações e padronizações, o que inclui corrigir erros, remover duplicatas, converter tipos de dados e aplicar regras de negócio básicas. Após essa transformação, aplicamos o Pydantic, uma biblioteca Python que valida o schema dos dados, ou seja, garante que todos os registros estejam consistentes com o formato

esperado. Isso assegura que apenas dados limpos e formatados corretamente sigam para as próximas camadas.

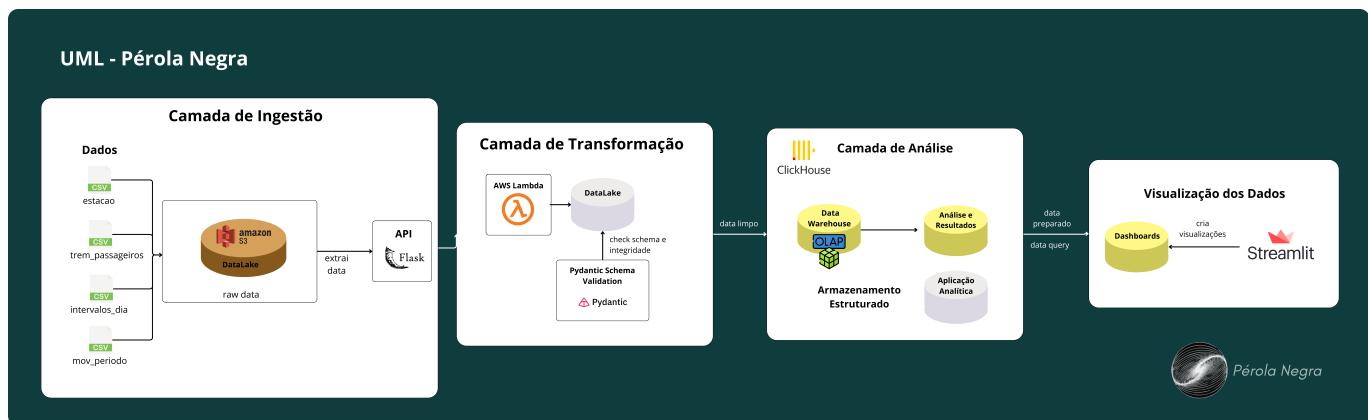
Camada de Ouro (Data Warehouse)

Na Camada de Ouro, os dados já transformados e validados na etapa anterior são agregados e otimizados para suportar análises mais avançadas e frequentes. Aqui, armazenamos esses dados em um Data Warehouse, no caso, o ClickHouse, que é um banco de dados analítico de alta performance. Na Camada de Ouro, os dados são preparados e organizados de forma a facilitar consultas rápidas e análises aprofundadas, sendo formatados para atender a consultas OLAP (Online Analytical Processing), que permitem operações complexas, como agregações e filtros eficientes.

Camada de Ródio (Visualização)

A Camada de Ródio foi criada para destacar a visualização dos dados, a etapa final e muitas vezes a mais valiosa para a tomada de decisões. Nessa camada, foi escolhido o Streamlit, hospedado em um servidor, para criar dashboards e infográficos interativos. Streamlit permite que visualizações dinâmicas sejam construídas de forma a transformar os dados processados em insights acionáveis. O grupo Pérola Negra usou o metal ródio para representar essa camada devido ao seu alto valor, simbolizando a importância e o impacto dos dados visualizados e analisados para o negócio.

Figura 6 - UML de Componentes - Pérola Negra



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Ingestão

A solução proposta para o pipeline de Big Data da CPTM começa com a Camada de Ingestão no AWS S3, onde dados brutos são armazenados diretamente no Data Lake, criando a base para a coleta de dados de várias fontes, como Estação, Trem_Passageiros, Intervalos_Dia, Mov_Período e Tipo_Embalagem. Essa coleta é feita por meio de uma API desenvolvida em Flask, que realiza a extração dos dados, enviando-os para o S3. A API é responsável por garantir que esses dados sejam capturados de forma precisa e segura, mantendo-os centralizados para o próximo estágio.

Transformação

Na transição para a Camada de Transformação, a solução utiliza o AWS Glue ou, alternativamente, AWS Lambda para executar o processo **ETL** (Extração, Transformação e Carga). Esta camada transforma os dados brutos em um formato mais utilizável, aplicando deduplicação, limpeza, conversão de tipos e normalização dos dados. Além disso, para garantir a qualidade dos dados, o Pydantic é usado para validar o schema e assegurar que os dados estejam no formato correto antes de avançarem no pipeline. Testes de consistência e integridade são executados dentro do diretório /test, verificando a conformidade e estabilidade dos dados transformados. Esse processo resulta em dados limpos e preparados, que são armazenados na Camada Prata do Data Lake.

Análise

Na Camada de Análise, o AWS EMR (Elastic MapReduce) com suporte de Apache Spark e Hadoop processa os dados transformados para análises distribuídas e cálculos estatísticos descritivos. Aqui, os dados da camada Prata são convertidos em informações significativas e análises valiosas, sendo preparados para consumo em dashboards e relatórios.

Visualização

Para a Visualização dos Dados, a ferramenta Streamlit é hospedada em um servidor, permitindo criar dashboards e infográficos intuitivos e visualmente atrativos para os usuários. Para visualizações avançadas, uma ferramenta open-source é utilizada em um container dedicado, facilitando a visualização interativa e dinâmica dos dados prontos para análise.

O Armazenamento Estruturado é feito em um Data Warehouse OLAP, armazenando os dados prontos para consultas analíticas, permitindo que a Aplicação Analítica acesse os dados para fornecer uma interface de interação com relatórios e infográficos finais, otimizando a gestão e tomada de decisões na CPTM.

3. Processo de ETL

ETL é a sigla para o processo de extrair, transformar e carregar. É uma forma tradicionalmente aceita para que as organizações combinem dados de vários sistemas em um único banco de dados, repositório de dados, armazenamento de dados ou data lake. O ETL pode ser usado para armazenar dados legados, ou, o que é mais comum, agregar dados para analisar e impulsionar as decisões de negócios. ([Google Cloud, 2024](#))

Por meio do ETL, é possível definir a qualidade dos dados e a forma como eles são manipulados a fim de transformá-los em uma informação inteligível e confiável.

O processo é composto por três etapas distintas:

- **Extração:** Consiste em coletar dados de diversas fontes, como bancos de dados, APIs, ou arquivos externos. O objetivo é centralizar as informações necessárias para análise em um único lugar, garantindo que sejam obtidos dados de qualidade e que as fontes de dados sejam confiáveis.
- **Transformação:** Nessa etapa, os dados são limpos, organizados e transformados para que possam ser utilizados de forma consistente. As transformações podem incluir a remoção de duplicatas, tratamento

de valores ausentes, normalização e agregação dos dados, tudo para que estejam prontos e adequados para o propósito de análise.

- **Carregamento:** Após a transformação, os dados são carregados no destino final, geralmente um Data Warehouse (OLAP), onde ficarão disponíveis para consultas e análises. É importante garantir que o processo de carregamento seja eficiente e que a integridade dos dados seja preservada ao longo do processo.

ETL

Arquitetura e Fluxo do Pipeline de Dados

Nesta seção, será apresentado o fluxo geral e as funcionalidades principais do código responsável pelo pipeline de ingestão de dados. Esse pipeline foi projetado para extrair dados armazenados em arquivos .parquet de um bucket S3, transformá-los e inseri-los em uma base ClickHouse, organizando e monitorando o processo para garantir consistência e rastreabilidade.

Estrutura Organizacional do Pipeline

O pipeline inicia com a estruturação de uma pasta denominada schemas, onde são definidos modelos de dados específicos para diferentes tipos de informações que serão processadas. Esses modelos – TrensPassageirosModel, IntervalosDiaModel, EstacaoModel, MovPeriodoModel e TipoEmbarqueModel – descrevem a estrutura de cada conjunto de dados e as respectivas especificidades, garantindo que todos os dados atendam aos requisitos do sistema final.

Função `get_parquet_files`: Identificação de Arquivos de Dados

A função `get_parquet_files` atua na identificação dos arquivos .parquet armazenados no bucket perola-negra. Ela realiza uma busca para localizar todos os arquivos relevantes que serão importados para ClickHouse. Esse processo inicial assegura que o pipeline tenha uma lista completa dos dados que precisam ser processados.

Função `convert_to_unix`: Padronização Temporal

Para manter a consistência dos dados, a função `convert_to_unix` transforma qualquer dado temporal em um formato Unix, facilitando a manipulação e interpretação. Esta etapa é essencial para garantir que todos os dados compartilhem uma linguagem temporal comum ao serem inseridos no ClickHouse, minimizando problemas de compatibilidade e processamento.

Função `read_parquet_and_insert_to_clickhouse`: Execução da Ingestão de Dados

Esta função gerencia o processo de leitura, transformação e inserção dos dados. Suas principais etapas são:

- *Criação de Tabelas:* A função inicia criando uma tabela no ClickHouse chamada grupo5.data_ingestion para armazenar os dados importados, com base nas estruturas previamente definidas.
- *Leitura e Preparação de Dados:* Cada arquivo .parquet é lido e convertido para o formato compatível com ClickHouse.

- *Inserção Condicional*: A função valida e insere os dados de acordo com o tipo (TrensPassageiros, IntervalosDia, etc.), garantindo que cada conjunto seja identificado corretamente no ClickHouse.
- *Validação e Tratamento de Erros*: O pipeline inclui um sistema de validação que emite alertas para dados inválidos ou com estrutura inconsistente, mantendo a integridade e a confiabilidade do pipeline.
- *Registro de Logs*: Todos os eventos, erros e sucessos são documentados em um sistema de observabilidade (log_observability), permitindo a auditoria e o acompanhamento das operações.

Função `ingest_data`: Coordenação Geral da Ingestão

Como principal orquestradora do pipeline, a função `ingest_data` executa a coleta, transformando e transferindo os dados para o ClickHouse. Ela percorre cada bucket identificado, processa os arquivos .parquet e, por fim, chama a função `read_parquet_and_insert_to_clickhouse` para realizar a operação de ingestão completa, com relatórios no sistema de logs.

Além disso, a lógica do pipeline considera chaves primárias e partições ao criar e carregar dados nas tabelas do ClickHouse, garantindo que a distribuição dos dados seja equilibrada e a consulta seja otimizada. Por exemplo, no caso da tabela `grupo5.data_ingestion`, o esquema definido no modelo `TrensPassageirosModel` inclui uma chave primária com base em identificadores do trem e timestamps, permitindo buscas rápidas e filtragem eficiente. Assim, ao carregar os dados, o pipeline verifica a conformidade das colunas com o esquema, garante o tipo correto de cada campo (ex: inteiro, string, datetime) e assegura que informações temporais sejam padronizadas, evitando divergências entre diferentes fontes. Caso algum registro não atenda aos padrões, o pipeline registra o evento e pula para o próximo lote, mantendo assim a consistência dos dados.

Resumo do Pipeline

- **Pasta schemas**: Definição e estruturação dos modelos de dados.
- **get_parquet_files**: Identificação dos arquivos .parquet a serem processados.
- **convert_to_unix**: Padronização dos dados temporais em formato Unix.
- **read_parquet_and_insert_to_clickhouse**: Função de ingestão com validação, estruturação e registro.
- **ingest_data**: Controladora geral que realiza a coleta, processamento e documentação dos dados.

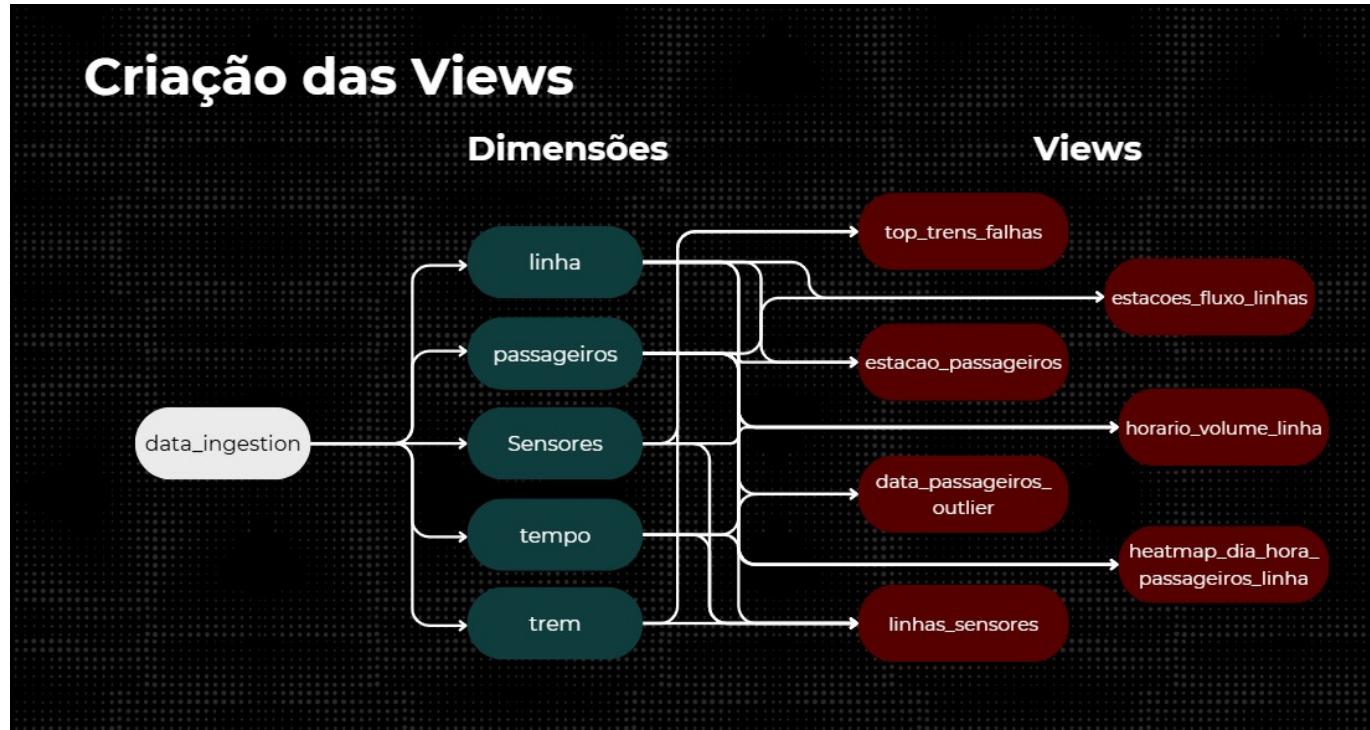
Este pipeline garante que cada conjunto de dados passe por uma estrutura organizada de extração, transformação e carga (ETL), com monitoramento e rastreabilidade, otimizando a integração com o ClickHouse e mantendo um histórico das operações realizadas.

3.1 Análise de Dados

Dimensões são estruturas fundamentais em um modelo de dados que organizam as informações em categorias ou grupos lógicos. No contexto deste projeto, as dimensões servem como bases estruturadas para agrupar, relacionar e acessar dados específicos de maneira eficiente. Elas permitem segmentar os dados em contextos relevantes para análises detalhadas e consultas direcionadas, facilitando a criação de relatórios e insights precisos.

As dimensões selecionadas para o projeto são fundamentais para a organização e análise dos dados no **DataApp**. Cada dimensão foi escolhida para garantir que as *views* criadas pudessem atender às necessidades operacionais e estratégicas da CPTM, melhorando a análise de dados relacionados ao transporte público. Abaixo é possível visualizar as dimensões e *views* utilizadas no projeto, seguidas de uma explicação detalhada.

Figura 7 - Dimensões e Views



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

1. **dimensao_estacoes:** Refere-se às estações da CPTM, incluindo detalhes como nome da estação, prefixos e atributos relevantes para análises. É essencial para mapear fluxos de passageiros e operações específicas por estação.
2. **dimensao_intervalo_dia:** Representa os intervalos diárias, como manhã, tarde e noite, sendo essencial para segmentar análises de operações ao longo do dia.
3. **dimensao_movimento_tempo:** Combina dados temporais com o movimento dos trens e dos passageiros. É crucial para análises de padrões operacionais e identificação de períodos críticos.
4. **dimensao_tipo_embarque:** Agrupa os tipos de embarque identificados, facilitando a análise do comportamento dos passageiros e das tendências de bilhetagem.
5. **dimensao_trens_passageiros:** Integra informações sobre os trens e os volumes de passageiros transportados, permitindo cruzamentos entre desempenho dos trens e ocupação.

As dimensões foram definidas levando em conta a lógica de chave estrangeira e chaves primárias para permitir junções eficientes entre tabelas. Por exemplo, a dimensão **dimensao_estacoes** possui um identificador único para cada estação (ex: `id_estacao`), que é utilizado nas views para relacionar eventos operacionais (como volume de passageiros ou falhas) a uma estação específica. Assim, ao consultar a view que mostra fluxo entre estações, o sistema realiza um join entre a tabela de fatos (com dados de fluxo) e a dimensão **dimensao_estacoes**, garantindo que os resultados sejam retornados rapidamente e de forma consistente. A normalização dessas dimensões segue padrões do tipo estrela (star schema), onde uma tabela fato central (como a de movimento de passageiros) se relaciona com dimensões que fornecem contexto. Esse design simplifica consultas OLAP, tornando a análise ágil e completa.

A criação de dimensões como tabelas facilita a manipulação e reutilização de dados estruturados, eliminando a necessidade de realizar extrações repetidas e otimizando o processamento de consultas.

As *views* criadas no **ClickHouse** organizam e sintetizam os dados coletados, sendo o alicerce para análises mais eficientes e direcionadas. Cada *view* foi desenvolvida para responder a questões operacionais específicas da CPTM, utilizando as dimensões previamente definidas.

1. **view_fluxo_entre_estacoes**: Relaciona o fluxo de passageiros entre diferentes estações. Baseia-se principalmente na dimensão **dimensao_estacoes** para entender os trajetos mais utilizados.
2. **view_heatmap_pessoas_por_linha**: Utiliza as dimensões **dimensao_estacoes** e **dimensao_movimento_tempo** para gerar mapas de calor com a distribuição de passageiros ao longo das linhas em períodos específicos.
3. **view_media_intervalo_operacao_por_dia**: Analisa a média dos intervalos de operação ao longo do dia, com base em dados temporais (**dimensao_intervalo_dia**).
4. **view_media_tempo_porta_aberta**: Apresenta o tempo médio em que as portas dos trens permanecem abertas, útil para otimizar a eficiência operacional. Relaciona-se com a dimensão **dimensao_movimento_tempo**.
5. **view_movimento_classificado_por_bilhete**: Classifica os movimentos dos passageiros com base nos tipos de bilhetes utilizados, utilizando a dimensão **dimensao_tipo_embarque**.
6. **view_sensores_por_data**: Relaciona dados coletados pelos sensores com a dimensão temporal, facilitando a análise de eventos ou falhas capturadas ao longo dos dias.
7. **view_tipos_bilhete_abundantes**: Identifica os tipos de bilhetes mais utilizados com base nos registros da dimensão **dimensao_tipo_embarque**.
8. **view_tipos_bilhete_por_dia**: Analisa o uso de bilhetes ao longo dos dias, cruzando dados temporais com os tipos de embarque.
9. **view_tipos_bilhete_por_semana**: Detalha o uso dos bilhetes durante a semana, permitindo identificar tendências sazonais e padrões de uso.

Cada *view* foi construída com base em consultas SQL otimizadas, que utilizam índices e partições presentes no ClickHouse. Por exemplo, em **view_fluxo_entre_estacoes**, a consulta utiliza filtragem por intervalos de tempo e junção com **dimensao_estacoes** para reduzir o conjunto de dados analisado, aumentando a velocidade da resposta. Com isso, a lógica interna da *view* garante que apenas as colunas necessárias sejam retornadas, diminuindo a carga no banco e melhorando a experiência do usuário final.

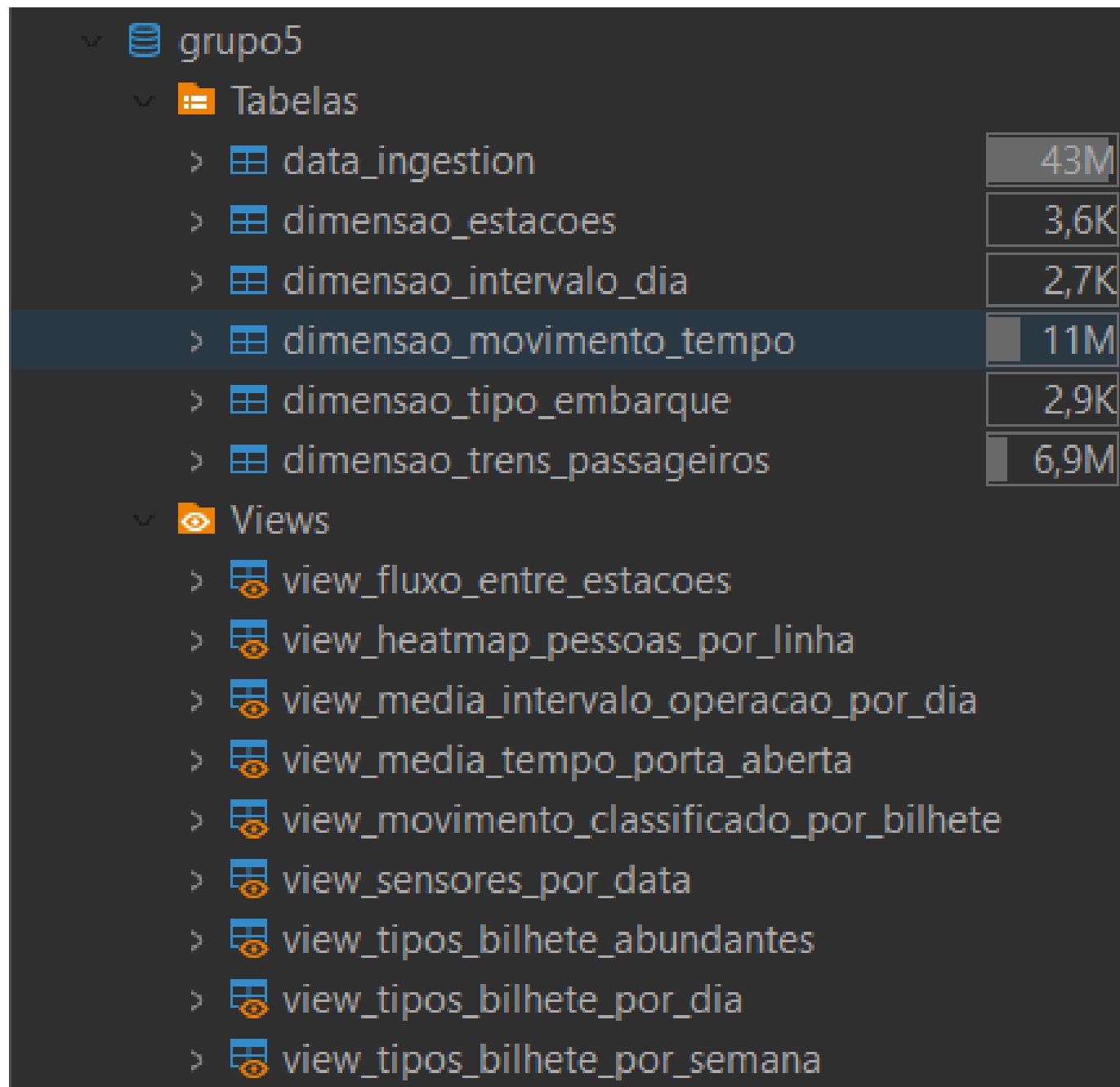
3.1.1 Criação das Planilhas

As tabelas de dimensões e *views* criadas no banco de dados **grupo5** foram desenvolvidas para organizar e estruturar os dados, permitindo análises alinhadas às necessidades operacionais. As dimensões mencionadas abordam aspectos essenciais do sistema, como:

- Estações e seus atributos (**dimensao_estacoes**)
- Intervalos temporais ao longo do dia (**dimensao_intervalo_dia**)
- Comportamento dos movimentos temporais (**dimensao_movimento_tempo**)
- Tipos de embarque identificados (**dimensao_tipo_embarque**)
- Dados específicos de trens e passageiros (**dimensao_trens_passageiros**)

Essas dimensões servem de base para cruzamentos e análises de dados mais específicas, permitindo insights para a operação da CPTM. Veja abaixo como está essa organização de forma mais visual.

Figura 8 - Dimensões e Views no DBeaver



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

A Figura acima apresenta a estrutura geral do banco de dados, com destaque para as dimensões e *views* que suportam as análises realizadas no projeto. Por exemplo, a tabela *dimensao_estacoes* armazena informações sobre as estações e as linhas associadas, incluindo o nome da estação, prefixos e descrição das linhas.

Figura 9 - Exemplo de dimensao_linha

The screenshot shows the ClickHouse Management Interface. On the left, the schema browser displays various tables and views under the 'dimensao_estacoes' database. On the right, a query results table for 'view_fluxo_entre_estacoes' is shown, with columns: id_estacao, tx_prefixo, tx_nome, and cd_estacao_bu|cluster. The table lists 13 rows of data.

	id_estacao	tx_prefixo	tx_nome	cd_estacao_bu cluster
1	ABR	ÁGUA BRANCA		511 2
1	ABR	ÁGUA BRANCA		511 2
4	AGN	ANTONIO GIANETTI NETO		709 0
4	AGN	ANTONIO GIANETTI NETO		709 0
5	AJO	ANTONIO JOÃO		558 2
5	AJO	ANTONIO JOÃO		558 2
6	ARC	ARACARÉ		753 0
6	ARC	ARACARÉ		753 0
8	BFI	BALTAZAR FIDELIS		502 2
8	BFI	BALTAZAR FIDELIS		502 2
9	BFU	PALMEIRAS-BARRA FUNDA		517 2
9	BFU	PALMEIRAS-BARRA FUNDA		517 2
10	BRU	BARUERI		557 2
10	BRU	BARUERI		557 2
11	BRR	BERRINI		606 2
11	BRR	BERRINI		606 2
12	BTJ	BOTUJURU		512 2
12	BTJ	BOTUJURU		512 2
13	BAS	BRÂS		764 0

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Além das dimensões, as *views* são elementos importantes na organização dos dados para análise. A *view* `view_fluxo_entre_estacoes`, por exemplo, destaca os fluxos com base em registros dos sensores. Ela permite identificar os problemas mais frequentes e priorizar ações de manutenção bom base no volume.

Figura 10 - Exemplo de View top_trens_falhas

The screenshot shows the ClickHouse Management Interface. On the left, the schema browser displays various tables and views under the 'dimensao_estacoes' database. On the right, a query results table for 'view_fluxo_entre_estacoes' is shown, with columns: estacao_inicio|estacao_fim|total_entradas|total_saídas. The table lists 6 rows of data.

	estacao_inicio estacao_fim	total_entradas	total_saídas
4	6	147543	224737
6	4	136324	122326
2	6	43673	23136
6	2	39319	74775
1	6	27712	11367
6	1	21464	41808

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Essas tabelas e *views* formam o alicerce do sistema de análise, fornecendo os dados que suportam a tomada de decisão operacional e estratégica que será exposta em infográficos. As dimensões fornecem a base estruturada, enquanto as *views* sintetizam os dados para análises específicas, como fluxo de passageiros, falhas de trens e desempenho operacional de linhas.

3.2 Cubo de Dados

O Prefect é uma plataforma de orquestração de workflows que permite a automação, monitoramento e gestão de tarefas complexas. Ele é amplamente utilizado para ETL (Extract, Transform, Load) e pipelines de dados, garantindo eficiência e controle em processos de atualização e análise. A principal vantagem do Prefect é sua capacidade de integrar diferentes fontes de dados e ferramentas, além de fornecer uma interface intuitiva para monitorar e gerenciar fluxos em tempo real.

No projeto da CPTM, o Prefect foi configurado para gerenciar a criação e atualização de todas as *views* no banco de dados ClickHouse. Essa configuração assegura que os dados sejam processados e organizados

automaticamente, eliminando a necessidade de intervenções manuais e garantindo que as informações estejam sempre atualizadas.

Cada view criada no ClickHouse está vinculada a um flow do Prefect, que encapsula as tarefas necessárias para sua criação ou atualização. Essas tarefas são configuradas para rodar em intervalos específicos ou sob demanda, dependendo da necessidade operacional.

A criação das views foi desenvolvida garantindo eficiência e clareza na organização das informações. Cada view é construída a partir de consultas SQL específicas, projetadas para transformar e estruturar os dados extraídos das tabelas de origem de forma consistente e otimizada. Como exemplo, abaixo está o código da view "top_trens_falhas".

```
from prefect import task
from config.connections import get_clickhouse_client
import os

@task(name="Create View Top Trens Falhas")
def create_top_trens_falhas_view():
    client = get_clickhouse_client()
    sql_query = """
        CREATE OR REPLACE VIEW grupo5.top_trens_falhas AS
        SELECT
            dt.train_id AS id_trem,
            ds.id_sensor AS id_sensor,
            ds.sensor_sts AS status_falha,
            COUNT(*) AS ocorrencias -- Conta as repetições
        FROM
            grupo5.dimensao_sensores AS ds
        JOIN
            grupo5.dimensao_trem AS dt ON ds.door_id = dt.door_id
        WHERE
            ds.sensor_sts != 0 -- Apenas falhas
        GROUP BY
            dt.train_id, ds.id_sensor, ds.sensor_sts
        ORDER BY
            ocorrencias DESC;
    """
    client.execute(sql_query)
    return "View 'top_trens_falhas' criada com sucesso!"
```

O código acima cria a view "top_trens_falhas" no ClickHouse usando Prefect para orquestração. A task `create_top_trens_falhas_view` executa uma consulta SQL que relaciona falhas dos sensores com os trens, filtrando apenas sensores com falhas, contando as ocorrências e organizando os resultados em ordem decrescente. A view é criada ou atualizada automaticamente, permitindo identificar os trens com maior número de falhas de forma eficiente.

Outras views seguem estruturas semelhantes, sendo possível de serem analisadas ao acessar o DBeaver de forma mais simples, como pode-se ver na imagem da “top_trens_falhas” a seguir.

Figura 11 - DBeaver - top_trens_falhas

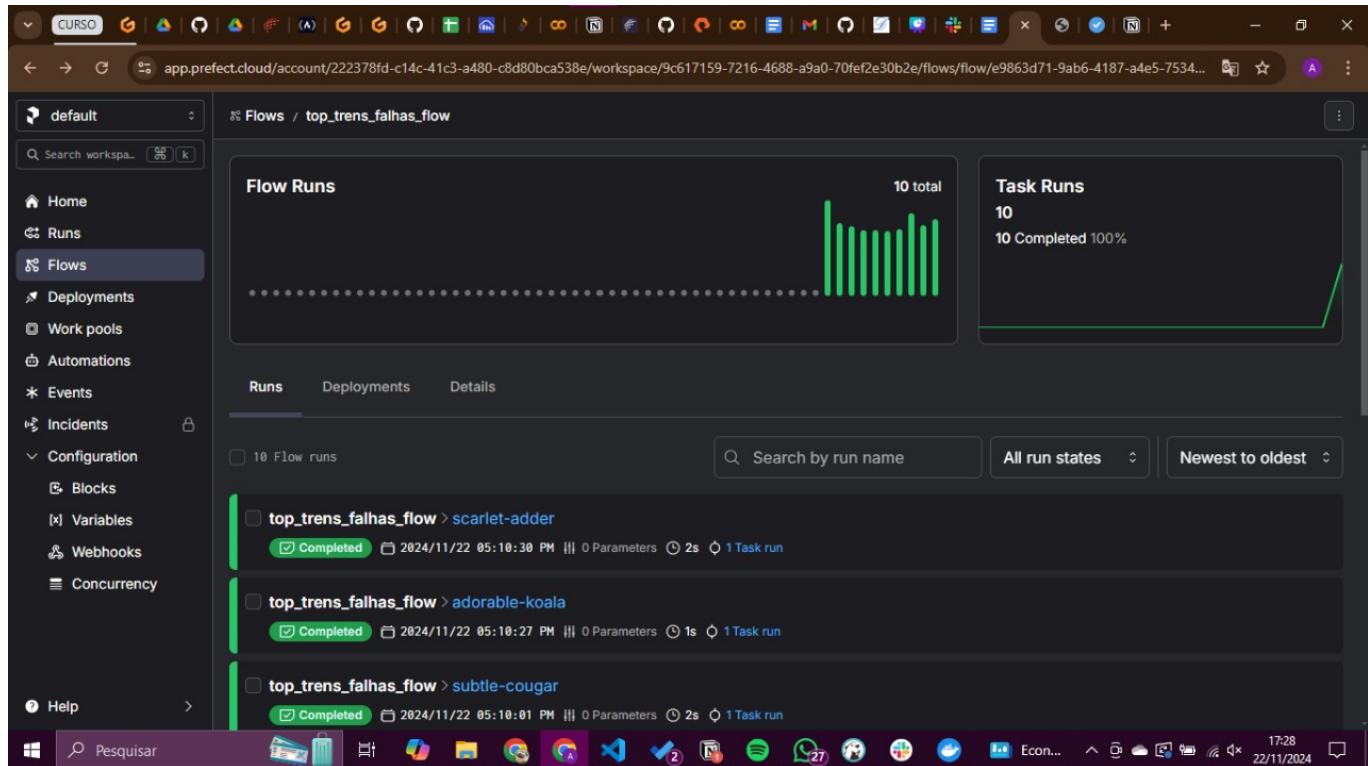
The screenshot shows the DBeaver 24.2.5 interface with the 'top_trens_falhas' view selected in the central workspace. The left sidebar displays the database structure, including tables like 'data_passageiros_outlier', 'linhas_sensores', and 'sensor_status_analysis'. The main area shows a table with the following data:

Grade	id_trem	id_sensor	status_falha	ocorrencias
1	-1	8	3	1.006.382
2	-1	16	3	928.968
3	-1	56	3	928.968
4	-1	48	3	851.554
5	-1	24	3	851.554
6	-1	4	3	851.554
7	-1	32	3	851.554
8	-1	64	3	851.554
9	-1	40	3	851.554
10	-1	80	3	774.140
11	-1	88	3	774.140
12	-1	96	3	774.140
13	-1	104	3	774.140
14	-1	72	3	774.140
15	-1	112	3	696.726
16	-1	120	3	696.726
17	-1	128	3	696.726
18	-1	12	3	696.726
19	-1	144	3	619.312
20	-1	176	3	619.312
21	-1	20	3	619.312
22	-1	208	3	619.312
23	-1	184	3	619.312
24	-1	152	3	619.312

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

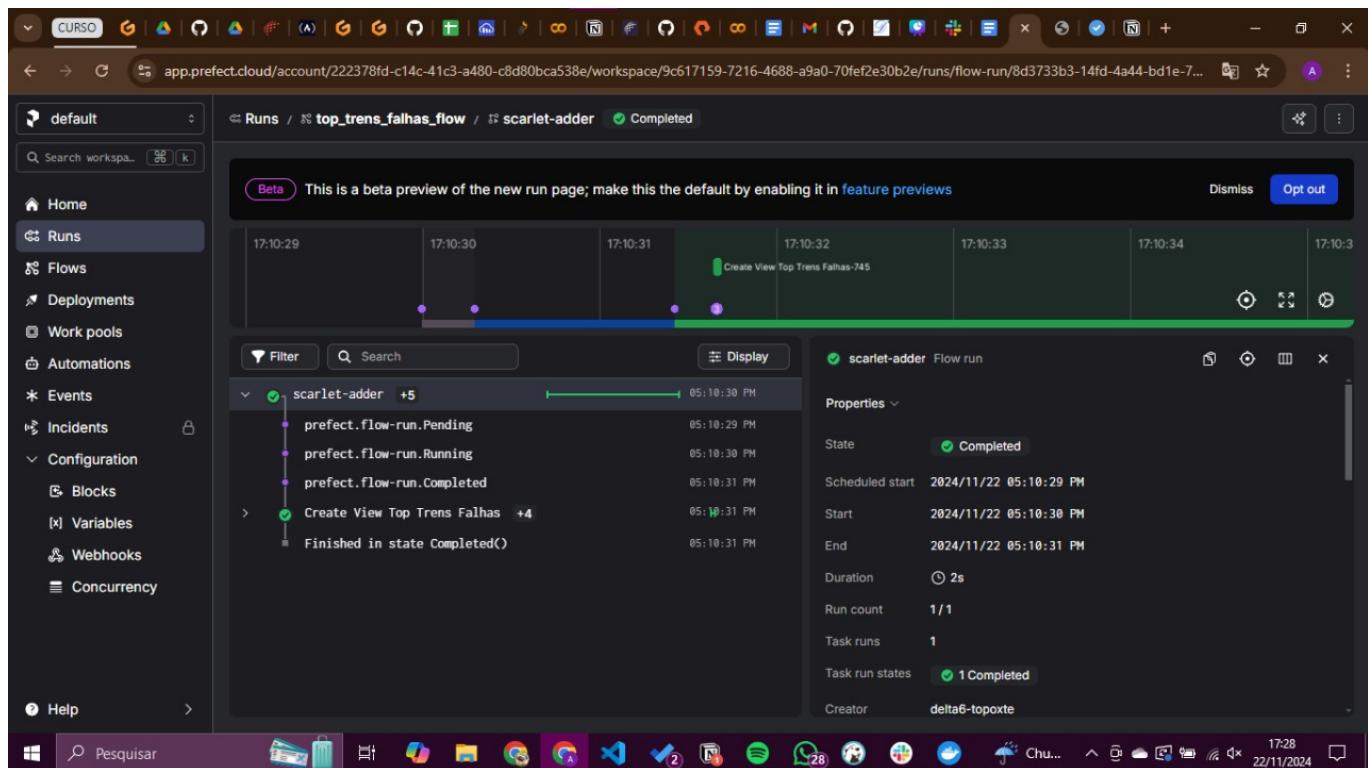
O Prefect oferece uma interface web que permite o monitoramento completo de todas as views configuradas no projeto. A plataforma registra logs detalhados de cada execução, incluindo informações sobre o status das tarefas, como erros ocorridos ou sucessos alcançados. Também é possível visualizar o estado atual dos flows, identificando quais já foram concluídos, quais estão em execução e quais aguardam na fila para serem processados. Em casos de falha, o Prefect facilita a reexecução manual ou automática do fluxo afetado, garantindo a continuidade do processo. Abaixo estão dois exemplos de visualização das views pelo Prefect.

Figura 12 - Histórico de Execuções Prefect



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Figura 13 - Monitoramento de Execução Prefect



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

A primeira imagem exibe a visão geral de todas as execuções do flow `top_trens_falhas_flow`. No painel, é possível ver o gráfico de barras que representa o histórico de execuções, destacando o número total (10) e a taxa de conclusão (100%). Abaixo, há uma lista com as execuções detalhadas, mostrando informações como o

nome atribuído automaticamente a cada execução (e.g., scarlet-adder, adorable-koala), o estado final (Completed), a duração (em segundos), e o número de tarefas executadas em cada flow. Essa visualização permite que a equipe acompanhe a consistência e o sucesso das atualizações da view ao longo do tempo.

Partindo para a segunda imagem, ela apresenta a interface de execução de um flow específico no Prefect, chamado top_trens_falhas_flow. Nela, é possível visualizar a linha do tempo da execução, indicando estados como Pending, Running e Completed, cada um com um marcador temporal. À direita, estão as propriedades do flow, que incluem o estado final (Completed), o horário agendado para início, o horário de término, e a duração total da execução (2 segundos). Essas informações são fundamentais para monitorar a eficiência e o desempenho das execuções. A lista de tarefas do flow, como Create View Top Trens Falhas, também é exibida, destacando que todas foram concluídas com sucesso.

Essa capacidade de monitoramento garante que a equipe possa identificar e resolver problemas rapidamente, minimizando impactos operacionais. Além disso, a automação reduz o risco de inconsistências nos dados, uma vez que todas as etapas do pipeline são rastreáveis e repetíveis.

4. Análise de Impacto Ético

Nessa seção é descrito os possíveis impactos de big data na sociedade e no meio ambiente.

Introdução

Essa seção visa oferecer uma análise de impacto ético envolvendo cinco dimensões, que são:

1. Privacidade e proteção de dados;
2. Equidade e justiça;
3. Transparência e consentimento informado;
4. Responsabilidade social;
5. Viés e discriminação.

Nesse contexto, é explorado essas cinco dimensões trazendo uma análise crítica de como o projeto pode impactar a sociedade para a CPTM. Diante disso, as palavras de Desmond Tutu, arcebispo sul-africano e ativista contra o apartheid, vêm para reforçar essa responsabilidade: "se você é neutro em situações de injustiça, você escolheu o lado do opressor."([Globo](#), 2024) Por isso, torna-se extremamente importante uma análise crítica, com recortes voltados para grupos minoritários na sociedade, para que realmente seja possível exercer um impacto positivo.

Impactos em Meio Ambiente e Sociedade

A Companhia Paulista de Trens Metropolitanos (CPTM) desempenha um papel social na mobilidade urbana no estado de São Paulo, atendendo diariamente a cerca de 1,6 milhão de passageiros ([CPTM](#), 2023). Nos aspectos ambientais, a CPTM, que declara utilizar energia limpa, sendo assim com o DataApp de Big Data, capaz de reduzir ainda mais as emissões e o consumo de recursos por meio de análises proporcionada pelos 5 grupos da turma de 2023 de Sistemas de Informação, otimizando o uso de energia no mínimo desperdício ([CPTM](#), 2023). A gestão de resíduos será também aprimorada, favorecendo a reciclagem e reutilização. Contudo, o impacto do armazenamento excessivo de dados requer atenção, pois dados

irrelevantes em data centers elevam o consumo de energia e, consequentemente, as emissões de carbono (Caetano, 2020). A eliminação de dados não é recomendada, uma vez que todos os dados operacionais e ambientais têm potencial valor para a CPTM. Contudo, é importante considerar o ciclo de utilidade dos dados, que pode durar aproximadamente três anos. Após esse período, dados que não apresentem mais relevância para operações atuais podem ser movidos para um arquivo de dados inativos, onde não haverá consumo significativo de processamento. Para otimizar o processamento, é essencial que a equipe da CPTM esteja capacitada para gerenciar dados de forma eficiente, evitando sobrecargas desnecessárias no sistema, uma vez que o processamento pode se tornar o maior problema, tanto por capacidade e por custo.

Em termos sociais, a inclusão e a acessibilidade serão beneficiadas por uma mobilidade urbana mais eficiente, com a redução do tempo de espera e maior frequência dos trens, promovendo a inclusão social (CPTM, 2022). A análise demográfica permitirá compreender as necessidades dos usuários, assegurando que os serviços atendam melhor as populações vulneráveis. Além disso, o uso de análise de Big Data permite identificar com mais clareza os locais e condições onde ocorrem incidentes, facilitando a implementação de medidas para mitigá-los e aumentando a segurança dos passageiros (CPTM, 2023). A transparência sobre operações e segurança pode fortalecer a confiança do público na CPTM.

Para garantir que o projeto de Big Data da CPTM esteja em sintonia com sua política ambiental e práticas ESG, é essencial adotar uma abordagem holística e integrada. Essa abordagem vai além da conformidade com legislações ambientais e inclui a consideração dos impactos sociais e ambientais gerados em toda a cadeia de operações da CPTM. A política de sustentabilidade da empresa já abrange ações em prol da redução de emissões, otimização de recursos naturais e iniciativas voltadas para o bem-estar das comunidades atendidas (CPTM, 2023). O uso responsável de Big Data nesse contexto pode fortalecer essa estratégia, permitindo uma análise mais detalhada e preditiva dos impactos ambientais, eficiência energética e melhorias operacionais. Dessa forma, o projeto se torna não apenas uma ferramenta de análise, mas um meio de suportar a missão ESG da CPTM, potencializando o compromisso da empresa com uma visão de futuro mais sustentável e integrada (CPTM, 2023).

A implantação de uma solução de Big Data na CPTM oferece uma oportunidade estratégica para o aprimoramento tanto ambiental quanto social. Ao focar em áreas-chave como eficiência energética, gestão consciente de dados e inclusão social, a CPTM poderá não apenas otimizar suas operações internas, mas também desempenhar um papel ativo na construção de um futuro sustentável. Para garantir o sucesso e a sustentabilidade desse processo, será fundamental estabelecer metas claras e viáveis, além de documentar e monitorar continuamente os resultados. Esse acompanhamento constante permitirá não apenas assegurar o cumprimento dos objetivos, mas também antecipar e mitigar possíveis impactos negativos. Para maior profundidade sobre a conformidade do projeto com o impacto ético, veja a análise das 5 dimensões a seguir: Privacidade e proteção de dados, Equidade e justiça, Transparência e consentimento informado, Responsabilidade social, Viés e discriminação.

4.1. Privacidade e Proteção de Dados

Essa seção documenta os métodos de coleta, armazenamento e utilização de dados no projeto, garantindo que as práticas adotadas estejam em conformidade com as regulamentações de proteção de dados e respeitem os princípios de privacidade.

Coleta de Dados

- **Métodos de Coleta:**

- Os dados são capturados através de **sensores** distribuídos nas estações e trens, além de **bases de dados externas** que contribuem com informações para a melhoria operacional. A coleta é justificada pela necessidade de melhorar a experiência do usuário, otimizar operações e aumentar a eficiência geral do sistema de transporte.
- Caso não sejam coletados dados que possam identificar indivíduos, deve-se registrar que a coleta é de **dados anônimos** ou agregados, que são caracterizados como **dados não pessoais**. Exemplos incluem dados como padrões de uso e fluxo geral de passageiros.

- **Tipos de Dados Coletados:**

- **Dados pessoais:** informações que poderiam identificar um indivíduo diretamente, como nome e CPF, só serão coletados se absolutamente necessários e mediante conformidade com a legislação.
- **Dados não pessoais:** incluem informações agregadas e não identificáveis, como o número de passageiros em intervalos específicos, padrões de utilização e informações sobre fluxo de transporte.
- **Transparência no processo de coleta:** todos os usuários serão informados sobre o uso de seus dados por meio de notificações claras e acessíveis. Isso inclui detalhes sobre a finalidade da coleta, o período de retenção e as medidas de segurança adotadas.

Armazenamento de Dados

- **Procedimentos de Armazenamento:**

- Os dados são armazenados na AWS S3, garantindo escalabilidade e alta disponibilidade para grandes volumes. Para garantir a flexibilidade, dados estruturados são salvos em CSV, enquanto dados semi-estruturados utilizam JSON, que facilita integrações futuras e análises avançadas. A arquitetura implementa o ClickHouse para consultas analíticas rápidas e eficientes, otimizando o processamento de grandes volumes de dados em tempo real. A ferramenta DBeaver está integrada ao ClickHouse, possibilitando a conexão direta para consulta e visualização dos dados, facilitando a exploração e análise de informações de maneira intuitiva. Para mais detalhes sobre a arquitetura e uso dos componentes, consulte a seção 2.2 UML de componentes.

- **Medidas de Segurança:**

- A segurança dos dados é garantida pelos recursos oferecidos pela **AWS**, que implementa proteção de dados de ponta para armazenamento, além de políticas rígidas de gerenciamento de identidades e permissões de acesso.
- **Controles de acesso rigorosos** são adotados, com restrições configuradas para que apenas pessoal autorizado tenha acesso aos dados sensíveis, utilizando autenticação segura e senhas fortes.
- Auditorias regulares e práticas de monitoramento contínuo são realizadas, assegurando a identificação rápida de tentativas de acesso não autorizado e promovendo a integridade dos dados.

Uso de Dados

- **Finalidades do Uso:**

- Os dados são empregados para fins específicos, como **melhoria nas operações de transporte**, otimização de horários, manutenção preventiva, e análise de desempenho dos serviços.
- Garantimos que os dados serão usados estritamente para os fins informados aos usuários, com relatórios regulares que asseguram a transparência no uso e divulgação.

- **Compartilhamento com Terceiros:**

- Caso ocorra compartilhamento de dados com terceiros, este será limitado a **parceiros estratégicos**, assegurando que estejam implementadas condições rigorosas para a proteção dos dados, sempre em conformidade com a legislação vigente.

Conformidade com a LGPD

- **Práticas de Conformidade:**

- Quando aplicável, será obtido o **consentimento explícito** dos usuários para a coleta e o uso de quaisquer dados pessoais.
- Foi implementado a política de **minimização de dados**, que assegura que somente os dados necessários ao projeto sejam coletados e mantidos.

- **Práticas de Proteção e Segurança:**

- São recomendadas auditorias e monitoramentos regulares para evitar acessos não autorizados, e todos os funcionários devem receber **treinamentos periódicos** sobre proteção de dados e práticas de segurança.
- Relatórios de transparência sobre o uso dos dados e medidas de segurança devem ser emitidos periodicamente por conta de CPTM, visando a conformidade com a LGPD e promovendo a confiança dos usuários.

A documentação deste processo deverá ser revista periodicamente, garantindo que todas as práticas se mantenham atualizadas em relação às regulamentações vigentes e às melhores práticas do setor, conforme indicado pela política de sustentabilidade da CPTM ([CPTM](#), 2023).

4.2. Equidade e Justiça

Objetivo: A equidade e a justiça social são princípios fundamentais que devem guiar qualquer iniciativa em uma empresa pública, especialmente em uma organização como a Companhia Paulista de Trens Metropolitanos (CPTM), cuja missão é fornecer serviços essenciais para a população. No contexto de um projeto de centralização de dados em um sistema de big data, é indispensável assegurar que o tratamento, análise e aplicação desses dados estejam alinhados com esses princípios. Isso inclui garantir que as informações coletadas correspondam à realidade e que não sejam perpetuados vieses que possam levar a decisões injustas ou à marginalização de determinados grupos.

A inclusão de dados que retratem de forma fiel as características e necessidades de diferentes segmentos da sociedade, como as pessoas com deficiência (PCD), é essencial para reduzir desigualdades. Dessa forma, significa estruturar o pipeline de dados para evitar preconceitos já existentes, como racismo e outras formas de discriminação, garantindo que os benefícios do projeto sejam realmente acessíveis a todos. Esse cuidado é indispensável para que o transporte público da CPTM seja mais justo, inclusivo e alinhado com os princípios de equidade que a empresa representa.

- **Identificação de Impactos:**

- Passageiros: Incluem-se aqui usuários regulares e específicos, como idosos, PCDs e pessoas em situação de vulnerabilidade social. O projeto pode afetar diretamente a experiência desses grupos ao influenciar decisões sobre infraestrutura, alocação de recursos e otimização de serviços.
- Equipes Operacionais: Com a centralização dos dados, mudanças nos processos de trabalho podem ocorrer, afetando a dinâmica e a capacitação de funcionários.
- Gestores e Planejadores: A análise centralizada permitirá decisões mais embasadas, mas pode também amplificar vieses se os dados não forem devidamente tratados.
- Barreiras de Acesso: Se os dados utilizados para o planejamento não representarem adequadamente a diversidade dos usuários, algumas populações, como as que residem em áreas periféricas ou possuem necessidades específicas, podem ser prejudicadas.

- **Estratégias de Mitigação:**

- Inclusão de Dados Diversos: Garantir que o banco de dados inclua informações detalhadas sobre todos os perfis de usuários, incluindo PCDs, idosos e grupos socioecononomicamente vulneráveis. A adição de dados fornecidos pela CPTM, especificamente relacionados a PCDs, é um exemplo positivo de como garantir a representatividade.
- Indicadores de Equidade: O grupo de passageiros da CPTM está liderando os esforços para monitorar e avaliar os indicadores diretamente relacionados à equidade. Isso inclui métricas como acessibilidade, frequência de trens e qualidade do atendimento em áreas com maior vulnerabilidade social.
- Auditorias Éticas e Técnicas: Realizar revisões periódicas no pipeline de dados para identificar e corrigir vieses que possam surgir na coleta, armazenamento e análise das informações.

4.3. Transparência e Consentimento Informado

Para assegurar que todas as partes interessadas no projeto de Big Data da CPTM tenham acesso claro e transparente às informações sobre o uso dos dados, garantindo que o consentimento seja obtido de forma informada e voluntária.

- **Comunicação com Usuários:**

- Fica sob responsabilidade da CPTM informar aos usuários e stakeholders da mesma sobre a coleta e o uso de dados por meio de **comunicados visuais, campanhas de conscientização e avisos em plataformas digitais** utilizadas pela CPTM. As informações são apresentadas de maneira clara e acessível, utilizando exemplos práticos, como cartazes explicativos nas estações, notificações nos aplicativos oficiais da CPTM e vídeos educativos em monitores dentro dos trens.
- Para garantir a atualização constante, também fica sobre responsabilidade da CPTM, as políticas de privacidade e os guias informativos, que devem ser revisados semestralmente e disponibilizados tanto em meios digitais quanto físicos, para fácil acesso por todas as partes envolvidas.
- Para reforçar o entendimento, as políticas de privacidade são traduzidas para linguagem simples e incluem exemplos cotidianos do impacto do uso de dados na melhoria dos serviços, como otimização de horários e manutenção preventiva.

- **Consentimento:**

- O consentimento é formalizado através de um **Termo de Consentimento Informado**, que especifica de maneira simplificada e visualmente acessível, detalhando quais dados serão coletados, suas finalidades e os direitos dos cidadãos sobre o uso e proteção de seus dados. Esse termo é elaborado em conformidade com a Lei Geral de Proteção de Dados (LGPD) e normas estaduais, reforçando o compromisso com a privacidade e os direitos dos cidadãos. Um exemplo desse documento pode ser encontrado na seção de Anexos, como o Anexo I.
- Todos os registros de consentimento são armazenados em um sistema seguro e auditável por responsabilidade da CPTM, garantindo que possam ser revisados futuramente para fins de conformidade e segurança. O processo de revogação de consentimento pode ser realizado diretamente por meio do aplicativo oficial da CPTM ou nas bilheterias físicas, garantindo que todos os cidadãos, independentemente de familiaridade com tecnologia, consigam gerenciar suas preferências de privacidade.
- Os dados coletados, sempre que possível, são apresentados de forma agregada e anonimizada, minimizando riscos de exposição indevida e priorizando a proteção dos indivíduos. Além disso, os cidadãos são informados previamente sobre qualquer mudança significativa nas práticas de coleta ou uso de dados por meio de e-mails, mensagens no aplicativo oficial ou comunicados em estações.
- A CPTM disponibiliza um canal de atendimento especializado para esclarecimentos sobre o consentimento e uso de dados. Esse serviço é oferecido por meio de plataformas de fácil acesso para que qualquer cidadão possa obter informações claras e tirar dúvidas, promovendo transparência e confiança no uso de dados em um serviço público ([CPTM. \(s.d.\)](#)).

Como afirmou o líder espiritual **Dalai Lama**, "A falta de transparência resulta em desconfiança e um profundo sentimento de insegurança" ([Lama, 2024](#)). Esse pensamento reforça a importância de práticas transparentes na coleta e uso de dados, que promovem a confiança entre a CPTM e seus stakeholders, essenciais para o sucesso do projeto.

A implementação dessas práticas não apenas a conformidade com a legislação aplicável, mas também promove uma comunicação efetiva e acessível, criando uma cultura de respeito à privacidade e à autonomia dos usuários, que são fundamentais em projetos de dados na era digital.

4.4. Responsabilidade Social

A ética dos dados refere-se aos princípios que orientam o uso responsável e justo das informações, assegurando que sua coleta, armazenamento, processamento e análise sejam realizados de forma transparente e em conformidade com padrões legais e morais. De acordo com o Gartner, trata-se de "um sistema de valores e princípios morais relacionados à coleta, ao uso e ao compartilhamento responsáveis de dados", com foco em todas as fases do ciclo de vida dos dados, desde sua geração até sua disseminação. No contexto de projetos baseados em Big Data, como o desenvolvido para a CPTM, a aplicação de padrões éticos é essencial para evitar prejuízos aos usuários e à sociedade, como a perpetuação de desigualdades e violações de privacidade. Assim, a ética deve ser incorporada desde a coleta dos dados até sua análise e apresentação.

4.5. Viés e Discriminação

O tratamento de dados em projetos de grande escala apresenta o risco de introduzir ou perpetuar vieses e discriminações. Esses problemas podem surgir de várias fontes, como dados históricos que carregam desigualdades sociais, processos de coleta enviesados, algoritmos com parâmetros inadequados ou até mesmo a falta de diversidade nas equipes responsáveis pelo desenvolvimento das soluções. Tais fatores, quando negligenciados, podem comprometer a justiça nas decisões, reforçar desigualdades e gerar impactos negativos para grupos sociais menos favorecidos.

Um exemplo prático de viés em um projeto como este seria a priorização de linhas com maior fluxo econômico, em detrimento de regiões periféricas que também enfrentam problemas críticos de transporte. Essa situação pode ocorrer quando os dados utilizados refletem apenas uma parte da realidade, deixando de lado as necessidades de áreas menos favorecidas. Além disso, falhas na modelagem de algoritmos podem resultar em análises que beneficiam desproporcionalmente certos grupos, intensificando a exclusão social.

Para mitigar esses problemas, algumas estratégias podem ser implementadas no projeto. Uma delas é a realização de auditorias frequentes nos dados, verificando sua representatividade em relação à diversidade de usuários do sistema ferroviário. É fundamental assegurar que o conjunto de dados reflita diferentes perfis demográficos, socioeconômicos e geográficos.

Outro aspecto crucial é o treinamento contínuo da equipe de desenvolvimento. É importante que todos os profissionais envolvidos tenham conhecimento sobre os riscos de viés e discriminação, além de compreenderem o impacto social das decisões baseadas em dados. Workshops e cursos sobre ética em ciência de dados, justiça algorítmica e design inclusivo podem ajudar a fortalecer essa perspectiva dentro do projeto.

Testes robustos também são indispensáveis para garantir que os algoritmos desenvolvidos funcionem de maneira justa em diferentes cenários. Por exemplo, ao definir prioridades para reparos em falhas ou alocação de recursos em horários de pico, os modelos devem ser avaliados considerando as demandas de diferentes grupos de passageiros, incluindo aqueles em situações de vulnerabilidade. Esse processo ajuda a evitar que decisões automatizadas beneficiem exclusivamente regiões ou populações específicas.

Por fim, a transparência em todo o processo de análise de dados é essencial para construir confiança. Relatórios claros e acessíveis devem ser disponibilizados, detalhando como os dados foram coletados, tratados e utilizados. Essa prática permite que stakeholders e a sociedade compreendam e questionem as escolhas feitas, promovendo responsabilidade e incentivando melhorias contínuas.

Em suma, abordar o viés e a discriminação em projetos de dados é uma tarefa que exige atenção técnica e compromisso ético. Ao adotar práticas inclusivas e ferramentas de detecção de viés, esse projeto pode garantir que as análises realizadas contribuam para decisões mais justas e impactem positivamente todos os usuários do sistema ferroviário. O combate a essas questões reforça não apenas a qualidade do trabalho, mas também seu alinhamento com os valores sociais de equidade e justiça.

4.6. Responsabilidade social

A responsabilidade social em projetos que utilizam Big Data ultrapassa as obrigações legais, incorporando um compromisso ético com a geração de impactos positivos para a sociedade. No contexto da CPTM, esse compromisso se traduz em tomar decisões informadas a partir dos dados coletados, promovendo melhorias diretas na qualidade de vida dos usuários e otimizando recursos em áreas que mais necessitam de atenção.

Um exemplo claro de responsabilidade social no projeto é a aplicação de análises para identificar horários e estações mais críticos. Essas informações permitem intervenções direcionadas, como aumentar a frequência de trens em momentos de pico ou melhorar a infraestrutura em estações com maior fluxo de passageiros. Além disso, o uso de tecnologias sustentáveis no armazenamento e processamento de dados reduz o impacto ambiental, alinhando o projeto aos princípios de desenvolvimento sustentável.

Outro aspecto relevante é o compromisso com a equidade. As decisões baseadas em dados devem priorizar a redução de desigualdades, direcionando recursos para áreas mais vulneráveis e garantindo que o sistema atenda de forma justa às diversas necessidades da população. Por exemplo, ao identificar regiões menos favorecidas com altos índices de demanda por transporte, a CPTM pode alocar mais trens ou implementar melhorias específicas, fortalecendo sua relação com a comunidade.

Além disso, o projeto contribui para a sociedade ao equilibrar eficiência operacional e impacto social. Ao implementar práticas que garantem decisões éticas e inclusivas, o projeto promove um transporte público mais justo, eficiente e sustentável. A coleta, análise e uso responsável dos dados fortalecem a credibilidade da CPTM e criam uma base sólida para o desenvolvimento contínuo.

Ao incorporar a responsabilidade social como um princípio norteador, o projeto vai além da análise de dados, posicionando-se como uma iniciativa que promove mudanças reais e duradouras. Essa abordagem assegura que os benefícios do projeto sejam compartilhados equitativamente, contribuindo para uma sociedade mais inclusiva e sustentável, enquanto reforça o nosso papel e o papel da CPTM como um exemplo de inovação ética e responsabilidade social.

Conclusão

A implantação do Big Data na CPTM representa um avanço importante, não só para melhorar a operação dos trens, mas também para fortalecer seu compromisso com sustentabilidade, inclusão e transparência. Ao trabalhar com indicadores éticos como privacidade, justiça, responsabilidade social e alinhamento à LGPD, o projeto vai além da tecnologia, focando no impacto positivo para a população e no uso consciente de recursos. Com metas claras e monitoramento constante, a CPTM reforça seu papel como referência em mobilidade urbana responsável, garantindo que os benefícios cheguem de forma justa a todos os usuários.

5. Streamlit e Infográfico

Essa seção é dedicada à documentação do que temos no nosso DataApp: o código **Streamlit** e o **Infográfico**, na última página do front-end.

5.1. Documentação do Streamlit

O Streamlit é uma biblioteca de código aberto em Python projetada para simplificar o desenvolvimento de aplicações web interativas e personalizáveis, voltadas para a visualização e exploração de dados. Criado em 2019, o Streamlit se destaca por sua abordagem intuitiva e minimalista, permitindo que cientistas de dados, analistas e desenvolvedores criem rapidamente interfaces gráficas sem necessidade de conhecimentos avançados em desenvolvimento web. ([Streamlit Documentation, 2024](#))

A principal vantagem do Streamlit é sua integração nativa com bibliotecas de ciência de dados populares, como pandas, NumPy, Matplotlib e Plotly, tornando-o uma escolha ideal para prototipagem rápida e compartilhamento de insights. Com comandos simples e um foco na produtividade, ele transforma scripts de Python em aplicativos web interativos executados localmente ou na nuvem em uma velocidade fora do normal.

O Streamlit foi utilizado como ferramenta para desenvolver um dashboard interativo que apresenta os dados da CPTM de forma visual e acessível. Com ele, os dados são processados e exibidos dinamicamente, permitindo que os usuários naveguem por diferentes páginas e explorem insights relacionados à operação ferroviária e ao comportamento dos passageiros.

5.1.1. Autenticação

Antes de o usuário chegar ao dashboard principal, é necessário passar por uma autenticação via login e senha. Esse mecanismo garante que apenas pessoas autorizadas acessem as informações operacionais e estratégicas, mantendo a confidencialidade dos dados e reforçando a credibilidade das análises. Ao fornecer suas credenciais, o usuário é identificado, o que facilita a rastreabilidade de suas ações, a auditoria interna e o monitoramento da utilização do sistema.

A experiência de login é simples e direta: ao acessar a plataforma, o usuário é imediatamente direcionado para a tela de autenticação, onde insere nome de usuário e senha. Caso as credenciais sejam válidas, a sessão é marcada como autenticada, liberando o acesso ao dashboard completo. Em caso de falha, uma mensagem orienta o usuário a verificar suas informações, assegurando assim que o acesso seja restrito apenas a indivíduos devidamente credenciados.

Na imagem abaixo, é possível visualizar a tela de login, que solicita ao usuário suas credenciais, garantindo que somente indivíduos autorizados acessem a solução.

Figura 14 - Tela de Autenticação



Fonte: Leandro Carvalho (2024)

Após o login bem-sucedido, o usuário é redirecionado ao dashboard principal, onde pode explorar métricas, relatórios e demais dados operacionais e estratégicos de forma segura, como pode ser visto na imagem abaixo:

Figura 15 - Template Data Product Canvas

A screenshot of a dark-themed dashboard titled "Visão Estratégica - CPTM". The sidebar on the left lists various operational metrics: Home, Fluxo Entre Estações, Heatmap, Intervalo Médio Operação, Tempo Porta Aberta, Movimento Classificado, Tipos de Bilhete, Sensores por Data, and Infográfico. A green notification bar at the top right says "Login bem-sucedido!". The main content area displays a title "Visão Estratégica - CPTM" with a small train icon, and a message "Bem-vindo ao painel estratégico. Explore os dados operacionais e filtre informações relevantes para a tomada de decisão." The top right corner shows deployment status: "RUNNING...", "Stop", "Deploy", and a zoom control "- 125% + Redefinir".

Fonte: Leandro Carvalho (2024)

Essas imagens ilustram o fluxo básico do usuário, desde o acesso inicial, passando pela autenticação, até a navegação no painel principal. Dessa forma, a camada de segurança não apenas protege a informação, mas também assegura uma experiência de uso estruturada e confiável.

5.1.2. Dashboard

A aplicação desenvolvida com Streamlit é estruturada em páginas, cada uma dedicada a um conjunto específico de análises e visualizações. Cada página obtém seus dados por meio de chamadas a endpoints GET, que interagem com a API Flask. Essa API atua como intermediária entre o dashboard e o banco de dados, garantindo segurança, atualizações constantes e integridade das informações apresentadas. Assim, o usuário tem sempre à disposição dados atualizados, pois o Streamlit executa as funções de coleta toda vez que uma página é carregada ou atualizada.

A estrutura das páginas é a seguinte:

Figura 16 - DataApp - Estrutura das Páginas



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

1. **Home:** Página inicial com uma introdução e navegação para as demais análises.
2. **Fluxo Entre Estações:** Exibe o fluxo de entrada e saída de passageiros em diferentes estações.
3. **Heatmap:** Apresenta um heatmap de movimentações de passageiros por linha e horário.
4. **Intervalo Médio Operação:** Mostra o intervalo médio de operação ao longo do dia.
5. **Tempo Porta Aberta:** Analisa o tempo médio em que as portas dos trens permanecem abertas.
6. **Movimento Classificado:** Classifica os movimentos por tipo de bilhete.
7. **Tipos de Bilhete:** Apresenta os tipos de bilhetes mais utilizados e sua distribuição ao longo do tempo.
8. **Infográfico:** Essa é uma página dedicada somente ao Infográfico desenvolvido em uma aula de UX, que foi explicado na próxima seção da documentação.

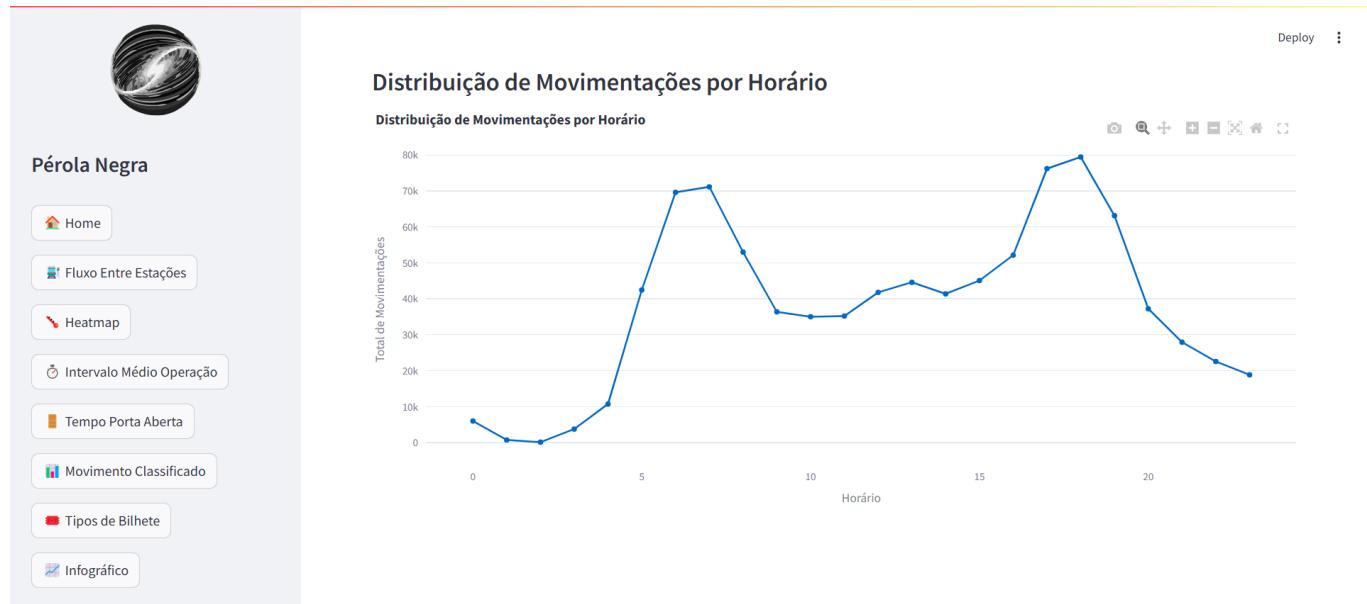
Toda vez que uma página é carregada, o Streamlit executa as funções associadas para realizar chamadas aos endpoints de GET. Isso significa que o dashboard está sempre atualizado com as informações mais recentes disponíveis no banco de dados.

As funções abaixo foram implementadas para coletar os dados necessários e alimentar cada página do dashboard:

- **get_fluxo_entre_estacoes:** Obtém o fluxo de passageiros entre estações.
- **get_heatmap_pessoas_por_linha:** Coleta dados para gerar um heatmap de movimentações por linha e horário.
- **get_media_intervalo_operacao_por_dia:** Recupera dados sobre o intervalo médio de operação durante o dia.
- **get_media_tempo_porta_aberta:** Calcula o tempo médio de porta aberta dos trens.
- **get_movimento_classificado_por_bilhete:** Classifica os movimentos dos passageiros por tipo de bilhete.
- **get_tipos_bilhete_abundantes:** Identifica os bilhetes mais utilizados.
- **get_tipos_bilhete_por_dia:** Analisa o uso de bilhetes por dia.
- **get_tipos_bilhete_por_semana:** Extrai dados semanais sobre os bilhetes utilizados.

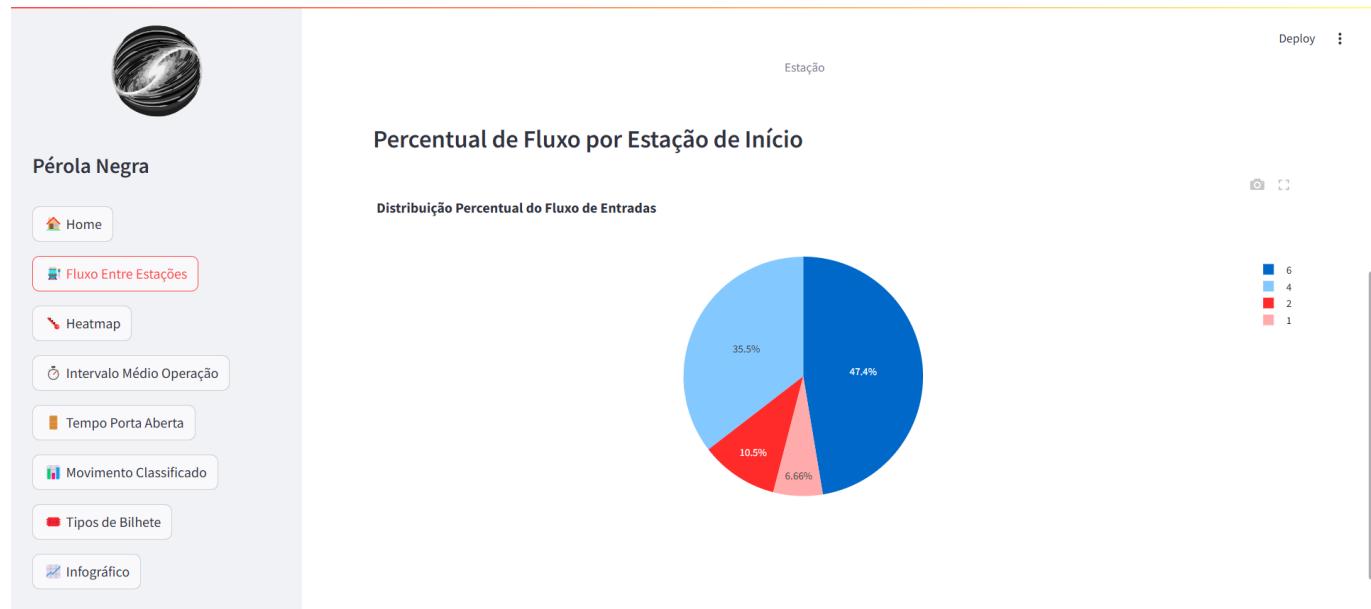
ANas imagens abaixo é possível visualizar como está a primeira versão do DataApp.

Figura 17 - DataApp - Distribuição de Movimentações por Horário



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Figura 18 - DataApp - Fluxo por Estação



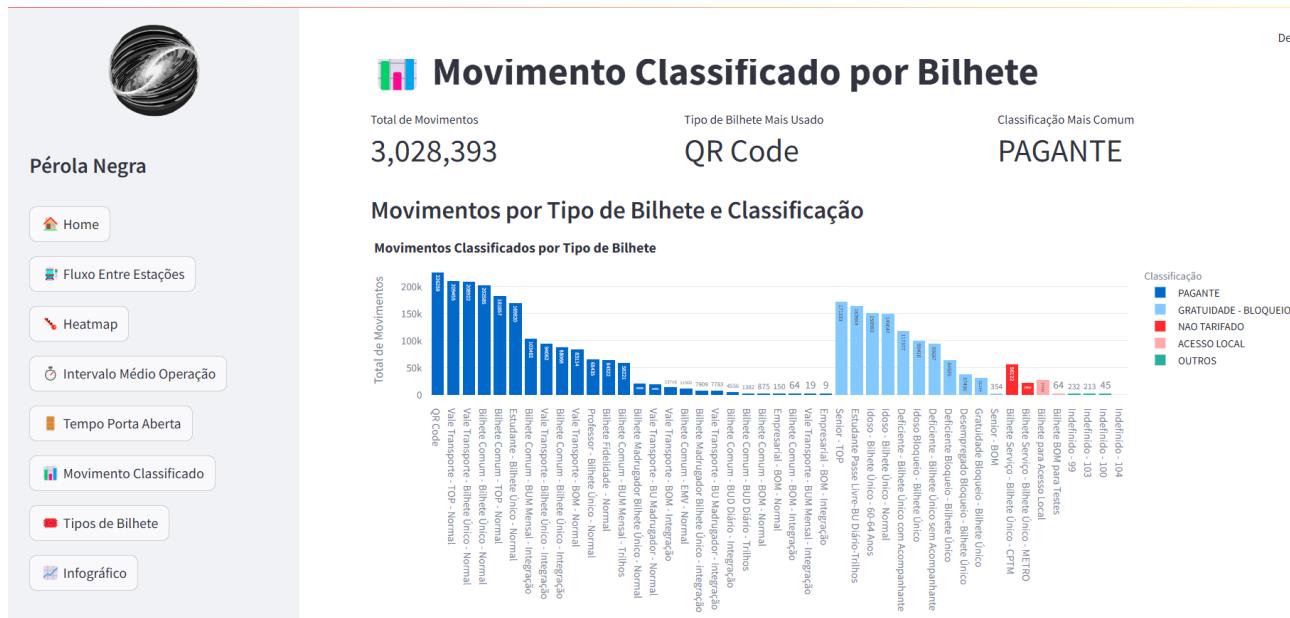
Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Figura 19 - DataApp - Tempo Médio de Porta Aberta



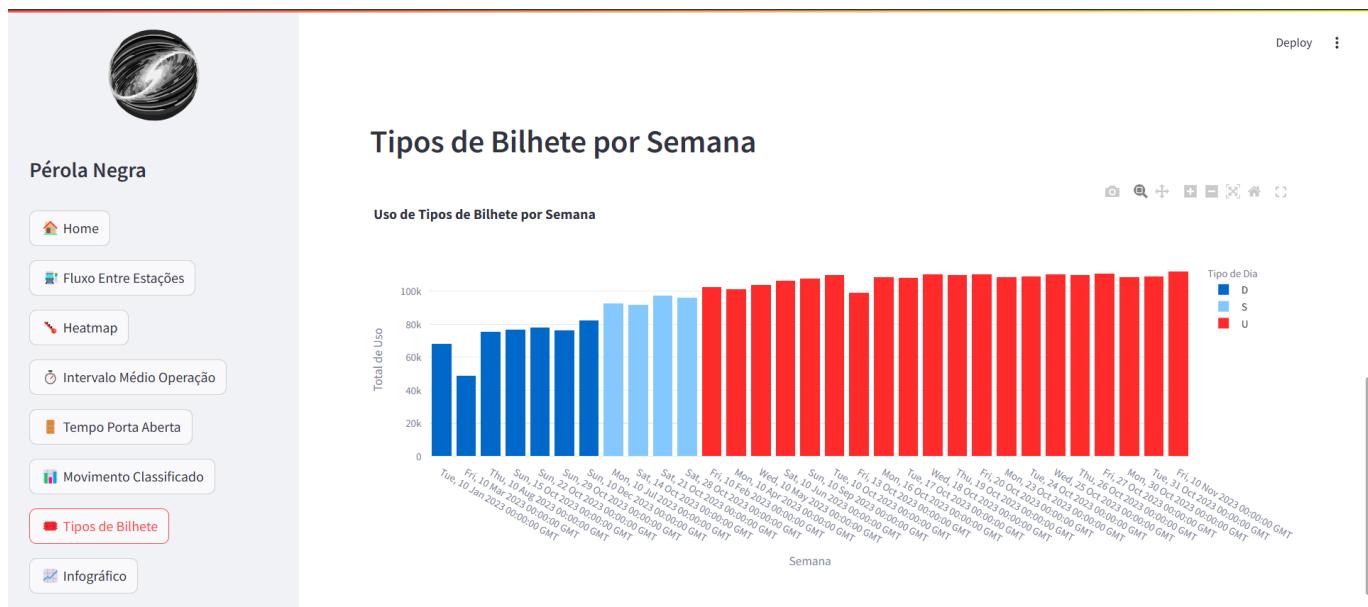
Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Figura 20 - DataApp - Movimento Classificado por Bilhete



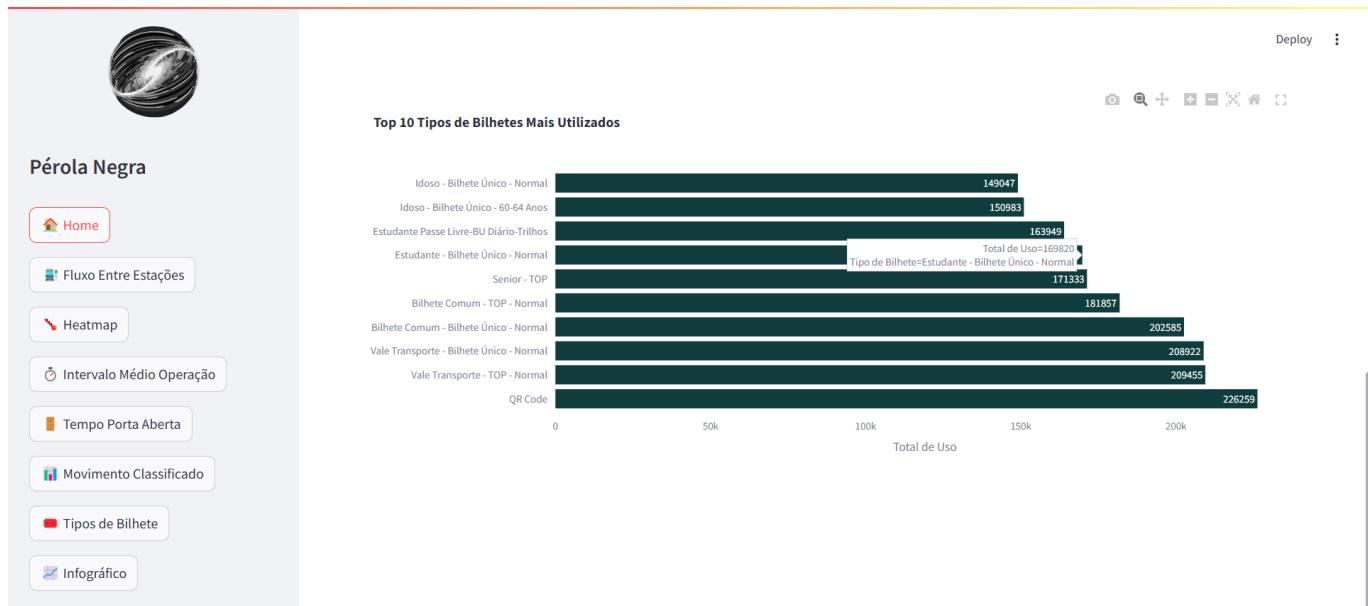
Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Figura 21 - DataApp - Tipos de Bilhete por Semana



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Figura 22 - DataApp - Top 10 Tipos de Bilhetes Mais Utilizados



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

O uso do Streamlit no projeto proporciona uma experiência interativa e dinâmica para a visualização de dados da CPTM. Com a atualização automática dos endpoints de GET a cada recarregamento de página, o dashboard garante que os dados exibidos estejam sempre atualizados, oferecendo uma base confiável para análises e tomadas de decisão.

5.2. Documentação dos Filtros

Além de todos os gráficos, também foram criados filtros para alguns deles, os quais estão descritos logo abaixo com imagens. Esses filtros servem para ajudar o usuário a encontrar mais rapidamente as informações específicas que procura.

Figura 23 - DataApp - Heatmap - Seleção de Linhas

Heatmap de Movimentação de Pessoas

Selecione a linha (ou 'Todas' para todas as linhas):

-
- Todas
- 13
- 97
- 98
- 99
- 100
- 2
- 3

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

No gráfico "Heatmap de Movimentação", o filtro de seleção de linha permite que o usuário escolha uma linha específica ou opte pela opção "Todas", incluindo assim todo o conjunto. Esse filtro traz flexibilidade ao analisar a movimentação dos passageiros, tornando possível focar em uma linha de interesse para entender padrões de uso, identificar horários de pico e necessidades operacionais relacionadas a determinadas rotas. Ao mesmo tempo, a opção de "Todas" facilita uma visão panorâmica de todas as linhas, auxiliando na comparação geral do desempenho e no planejamento estratégico.

Figura 24 - DataApp - Filtro - Heatmap Intervalo de Horários

Heatmap de Movimentação de

Selecione a linha (ou 'Todas' para todas as linhas):

13

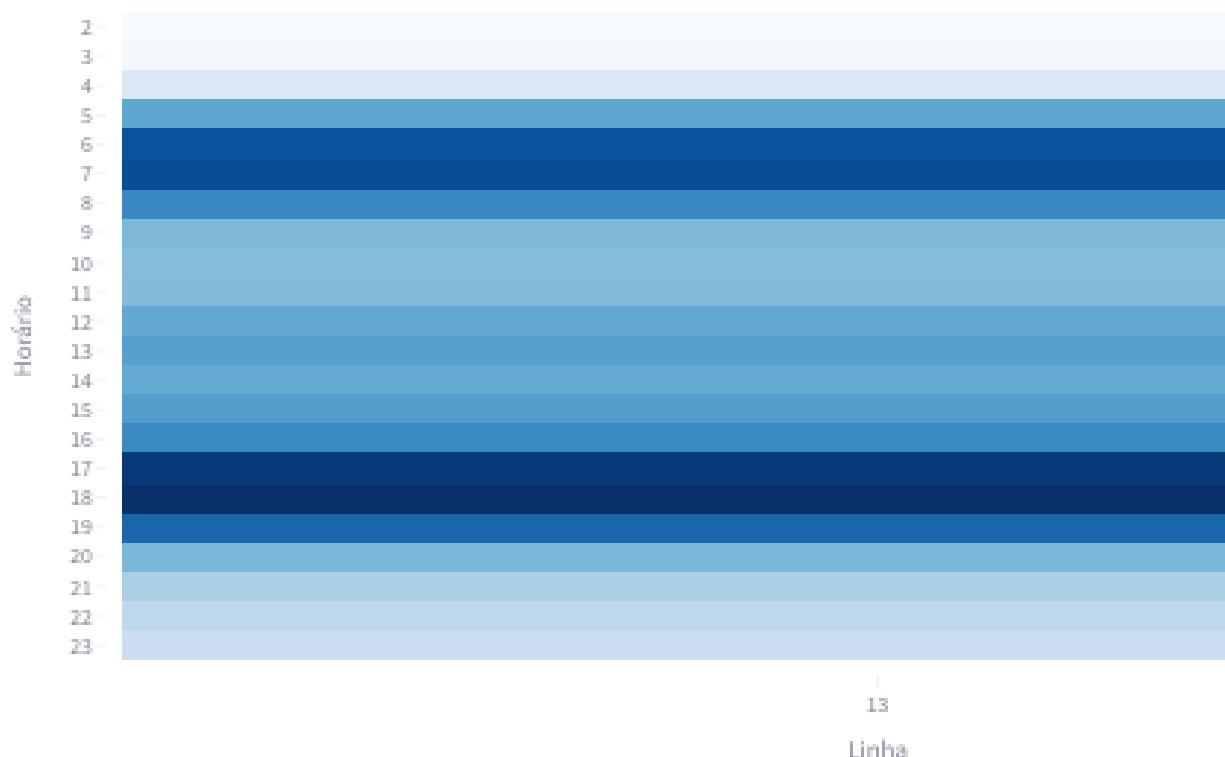
Selecione o intervalo de horários:

0 2



Mapa de Calor - Movimentação por Linha e Horário

Heatmap - Movi



Total Movimentações

890,773 pessoas

Linha Mais Movimentada

Linha 13

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Após selecionar a linha, como mostrado na figura 24, o filtro torna possível a seleção de intervalo de horário, por meio de uma barra vermelha deslizante. Esse recurso permite que o usuário limite a análise a um período específico do dia, seja o horário de pico da manhã, o final da tarde ou um intervalo pré-determinado. Ao ajustar o intervalo, o usuário pode detectar padrões temporais de movimentação, identificar gargalos nos horários de maior demanda e ajustar recursos, como número de trens ou funcionários, para melhorar a eficiência operacional.

Figura 25 - DataApp - Filtro - Eventos Críticos por Data

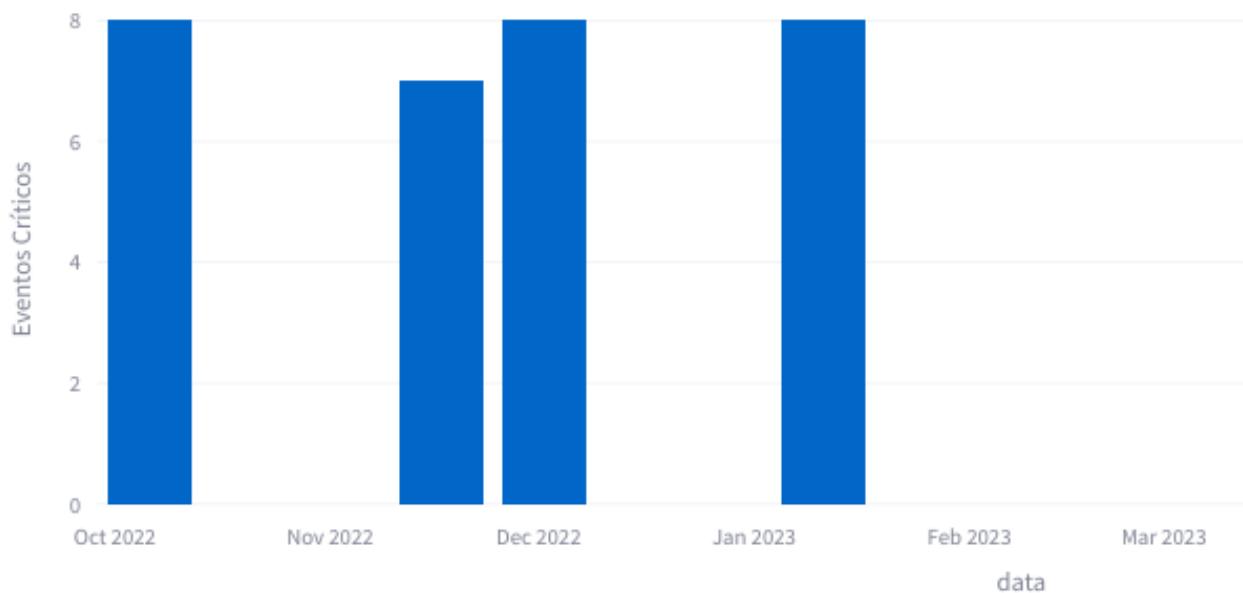
Eventos Críticos por Data

Intervalo de datas para o Gráfico 4

2020-09-17



Eventos Críticos por Data



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

No gráfico "Eventos Críticos por Data", o filtro de intervalo de datas possibilita que o usuário selecione um período específico para análise. Essa funcionalidade é essencial para investigações temporais, permitindo observar se houve aumento de eventos críticos em certos meses, ou a queda após iniciativas de manutenção. Ao restringir o período, é mais fácil correlacionar os incidentes com fatores externos (como clima, feriados ou eventos na cidade) e tomar decisões informadas sobre alocações de recursos e ações preventivas.

Figura 26 - DataApp - Filtro - Tendência de Eventos ao Longo do Tempo

Tendência de Eventos ao Longo do Tempo

Intervalo de datas para o Gráfico 3

2020-09-17

2020-09-17

Tendência de Eventos por Status



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

No gráfico “Tendência de Eventos ao Longo do Tempo”, o filtro de datas atua como um zoom temporal, permitindo avaliar tendências e flutuações em períodos específicos. Ajustando esse intervalo, o usuário pode analisar a evolução dos eventos ao longo dos anos, verificar se as medidas corretivas implementadas surtiram efeito e identificar padrões de sazonalidade. Essa visão refinada apoia o planejamento de longo prazo e a melhoria contínua do serviço.

Figura 27 - DataApp - Filtro - Proporção de Status dos Sensores

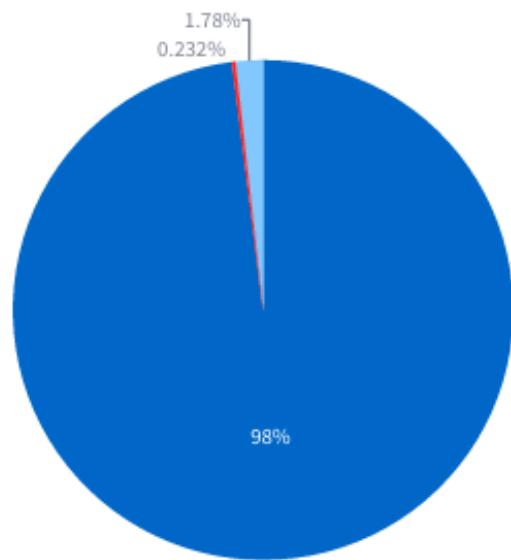
Proporção de Status dos Sensores

Intervalo de datas para o Gráfico 2

2020-09-17

2020-09-17

Distribuição Percentual dos Status



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

No gráfico “Proporção de Status dos Sensores”, o filtro de datas delimita a janela de observação para analisar a condição dos sensores em determinado intervalo. Desse modo, é possível verificar se houve variações significativas nas proporções de status em períodos específicos. Essa informação é útil para monitorar a eficácia da manutenção preventiva, avaliar a estabilidade do sistema de sensores e antecipar intervenções técnicas antes que um volume maior de falhas ocorra.

Figura 28 - DataApp - Filtro - Total de Eventos por Data por Linha

Total de Eventos por Data (Linha Selecionada)

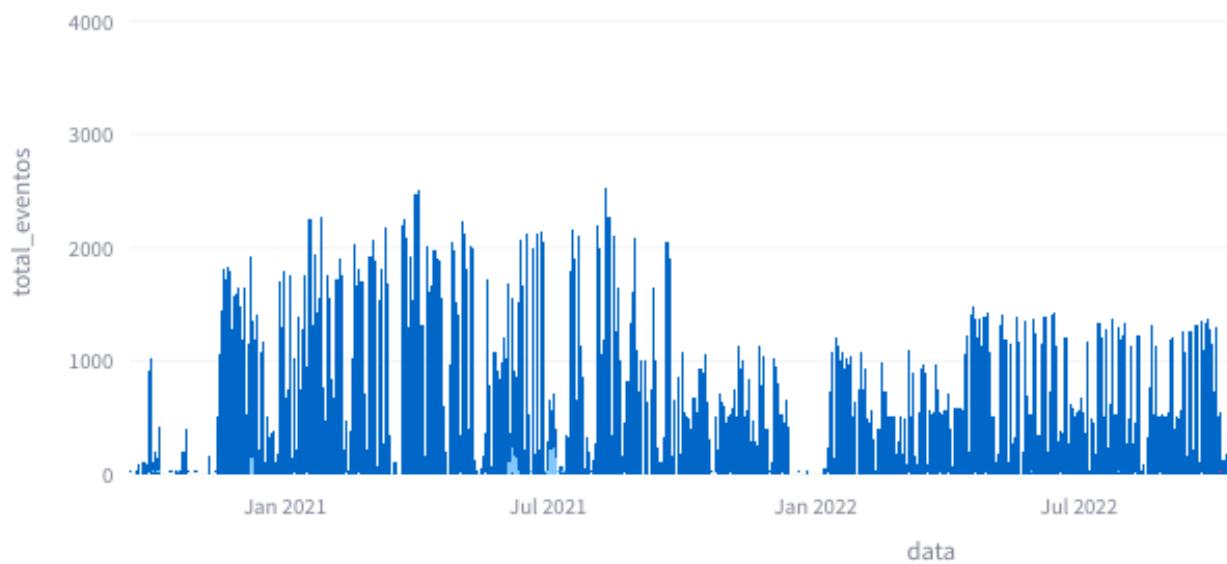
Intervalo de datas para o Gráfico 1

2020-09-17



2020-09-17

Eventos por Status - Linha Todos



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

No gráfico “Total de Eventos por Data (Linha Selecionada)”, o filtro de datas combinado à seleção de linha permite entender a dinâmica de eventos ao longo do tempo em um contexto mais restrito. Ajustando o intervalo de datas, o usuário pode analisar períodos críticos, como semanas de manutenção intensiva ou temporadas com demanda atípica, e avaliar o impacto em linhas específicas. Essa abordagem facilita a correção de falhas, a otimização de rotas e a melhoria na alocação de recursos.

Figura 29 - DataApp - Filtro - Sensores por Data

Sensores por Data

Selecione uma Linha:

Todos

OK

Warning

3800

69

Total de Eventos por Data (Linha Selecionada)

Intervalo de datas para o Gráfico 1

2020-09-17

2020-09-17

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

No painel “Sensores por Data”, o filtro de seleção de linha permite que o usuário escolha uma linha específica ou mantenha a opção “Todos”. Ao focar em uma linha específica, é possível detectar problemas pontuais de sensores naquele trajeto, entender padrões de falhas e necessidades de manutenção mais frequentes. Já a opção “Todos” oferece uma visão geral, possibilitando a comparação entre diferentes linhas e auxiliando na priorização de investimentos em tecnologia e manutenção.

Figura 30 - DataApp - Filtro - Entradas e Saídas por Estação

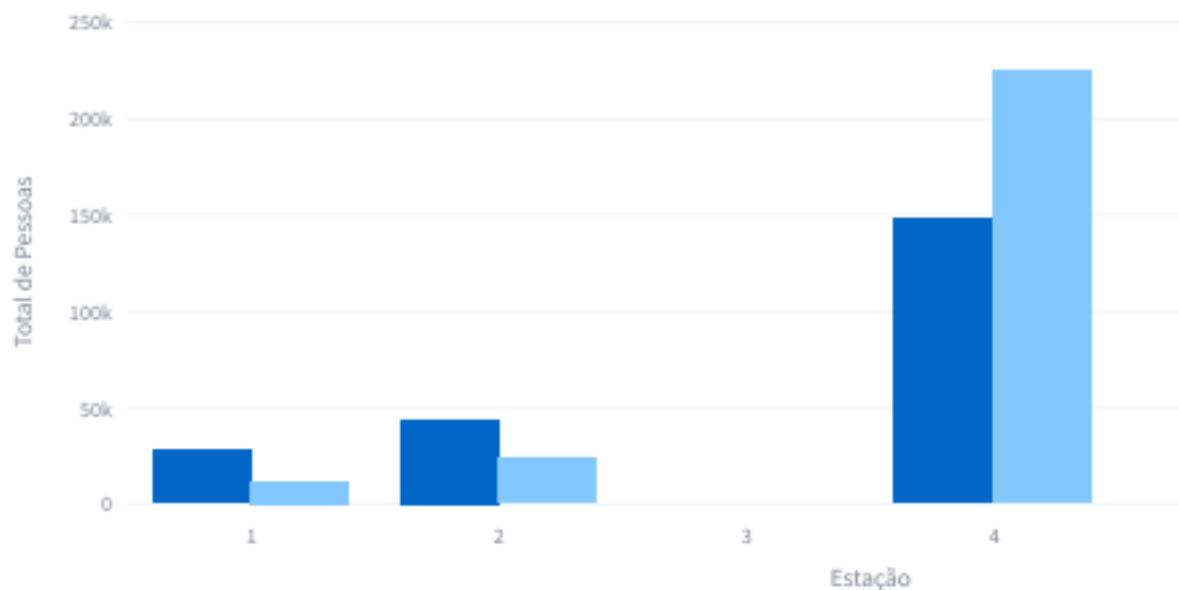
Fluxo Entre Estações

Selecione as estações:

Todos ×

Gráfico de Entradas e Saídas por Estação

Comparativo de Entradas e Saídas em todas as datas



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

O filtro acima, utilizado no gráfico "Fluxo Entre Estações," é responsável por permitir que o usuário escolha quais estações deseja visualizar, além de incluir opções para selecionar todas ou nenhuma estação. Esse filtro é essencial para personalizar a análise, permitindo foco em estações específicas de interesse ou uma visão ampla do fluxo entre todas as estações. Ele facilita a identificação de padrões ou anomalias em determinadas estações, auxiliando na tomada de decisões baseadas em dados operacionais.

Figura 31 - DataApp - Filtro - Tempo Médio de Porta Aberta por Linha

Tempo Médio de Porta Aberta

Data de início:

2020/09/17

Data de fim:

2023/09/23

Resumo do Tempo Médio de Porta Aberta por Linha

Linha 13.0 ⓘ

13.62 min

↑ 10.11 min

Linha 97.0 ⓘ

3.65 min

↑ 20.08 min

Linha 98.0 ⓘ

48.64 min

↑ 24.91 min

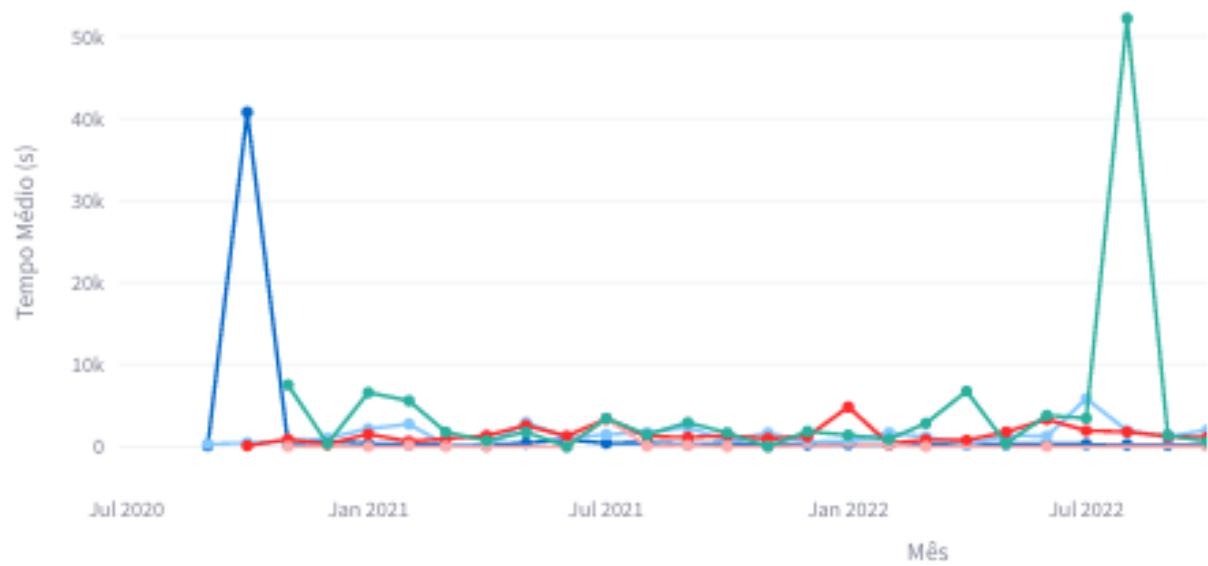
Linha

25

↑ 1

Variação Mensal do Tempo Médio de Porta Aberta por Linha

Variação Mensal do Tempo Médio de Porta Aberta



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

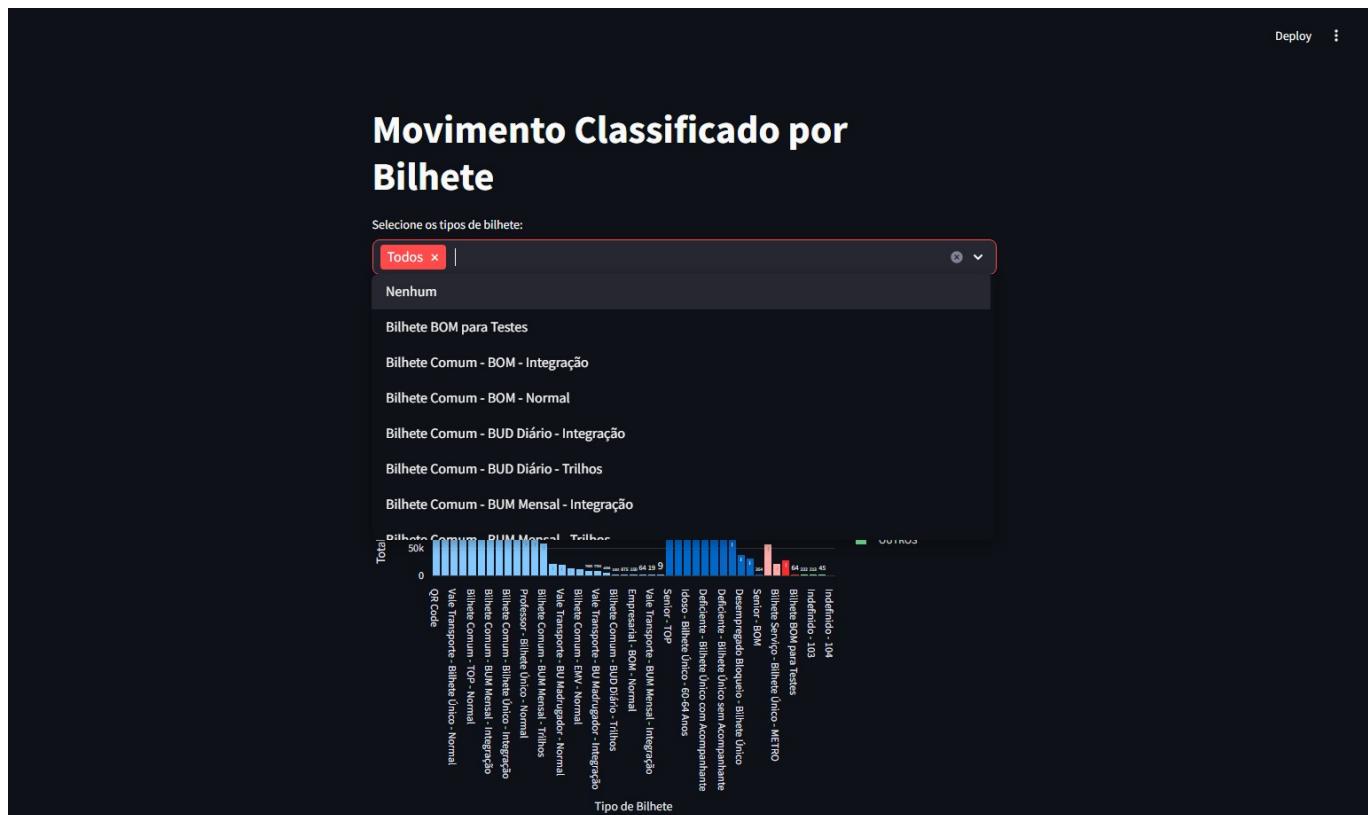
Na visualização do "Tempo Médio de Porta Aberta", o usuário pode ajustar as datas de início e fim, delimitando o período a ser analisado. Dessa forma, é possível observar se o tempo médio de porta aberta aumentou ou diminuiu após certa intervenção, se determinados meses são mais críticos devido ao clima ou horários de pico, e quais linhas apresentam maior variação ao longo de um intervalo. Esse recurso garante uma análise mais precisa e contextualizada, auxiliando no aprimoramento da operação.

Figura 32 - DataApp - Filtro - Movimento Classificado por Bilhete 1



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Figura 33 - DataApp - Filtro - Movimento Classificado por Bilhete 2

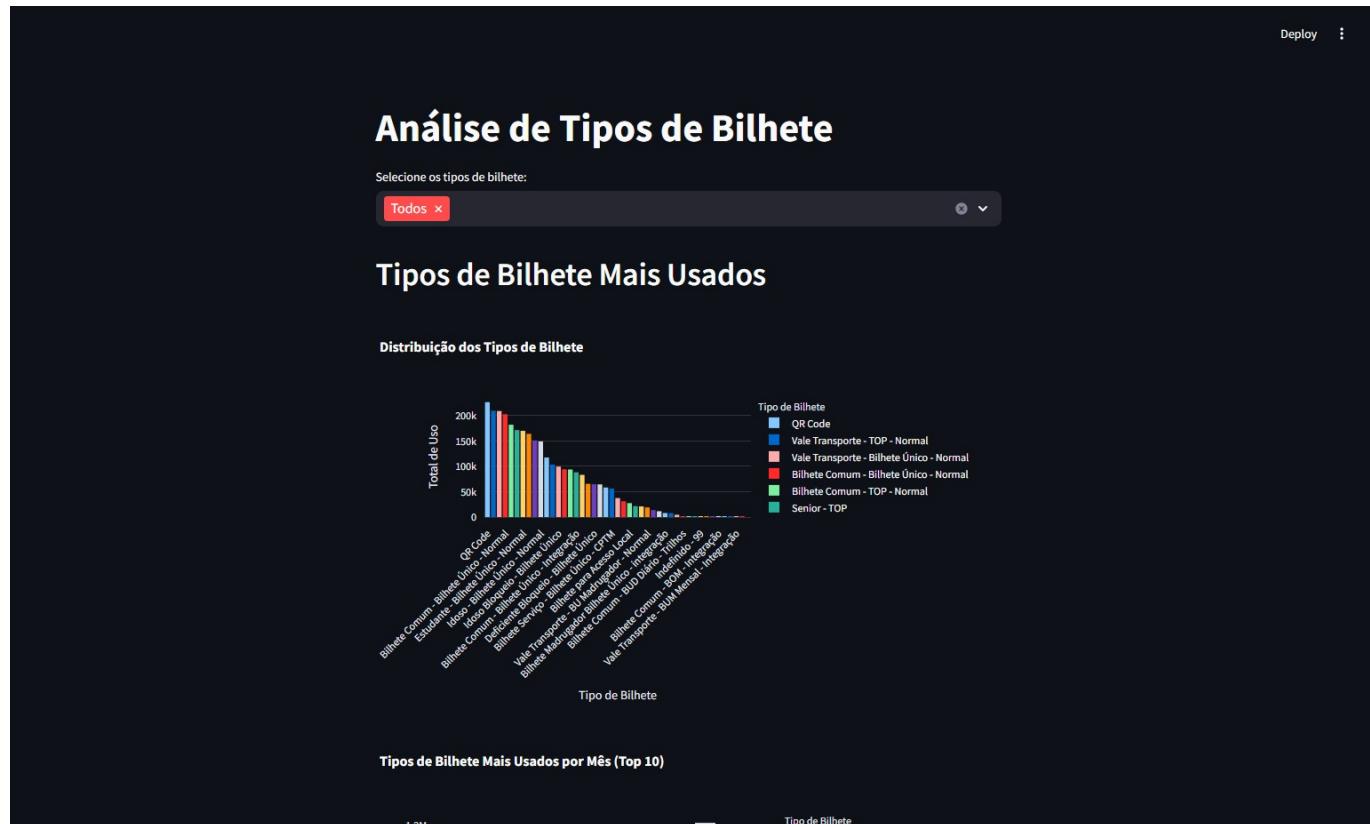


Fonte: Material produzido pelo Grupo Pérola Negra (2024)

O filtro acima, do gráfico "Movimento Classificado por Bilhete", tem uma função parecida com o primeiro filtro. Ele permite que o usuário selecione o bilhete que deseja visualizar, além de oferecer as opções de exibir

todos ou nenhum bilhete. Esse filtro é importante para análises específicas por tipo de bilhete, ajudando a identificar tendências no uso de categorias específicas, como bilhetes de estudante, VT ou QR Code. Além disso, o gráfico associado oferece a funcionalidade de zoom, permitindo que o usuário amplie partes específicas, como dias da semana, para uma análise mais detalhada e segmentada, fornecendo insights valiosos para a otimização do sistema de bilhetagem e o atendimento aos passageiros.

Figura 34 - DataApp - Filtro - Análise de Tipos de Bilhete



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Por fim, o filtro acima (de Análise de Tipos de Bilhete) é extremamente parecido com o anterior, pois também faz uma seleção de bilhetes. Porém, ele se limita a mostrar um gráfico de barras com os tipos de bilhetes mais utilizados, sem considerar dados específicos por dia da semana ou períodos, dado que é mostrado em outro gráfico. Essa simplicidade torna o filtro eficiente para identificar rapidamente os bilhetes mais populares, auxiliando na priorização de estratégias focadas nas categorias mais utilizadas, como melhorias no atendimento ou campanhas promocionais direcionadas.

5.3. Documentação do Infográfico

Como foi dito na seção anterior relacionada ao código do Streamlit, a última página do DataApp foi dedicada para mostrar um infográfico. Ele foi criado para apresentar uma retrospectiva histórica e destacar dados relevantes sobre o uso dos trens e bilhetes da CPTM. Ele combina um storytelling visual sobre a evolução dos trilhos e os tipos de bilhetes mais utilizados atualmente, permitindo que os usuários explorem essas informações de forma interativa. A ideia principal é conectar o impacto histórico dos trens ao comportamento contemporâneo de seus usuários.

Abaixo pode-se conferir uma imagem do mesmo.

Figura 35 - Infográfico

Dos Trilhos à Rotina

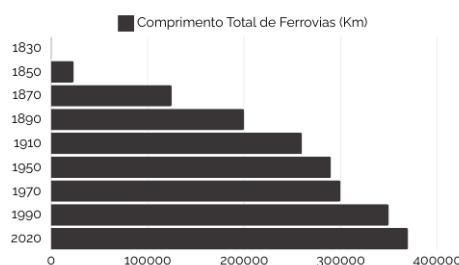
A Evolução dos Trens como Meio Essencial para Trabalhadores

A Revolução Industrial e os Trilhos

"A História na Europa"

Século XIX

No século XIX, as ferrovias impulsionaram a Revolução Industrial na Europa, facilitando o transporte de matérias-primas, mercadorias e trabalhadores para as fábricas nos centros urbanos.



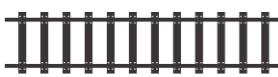
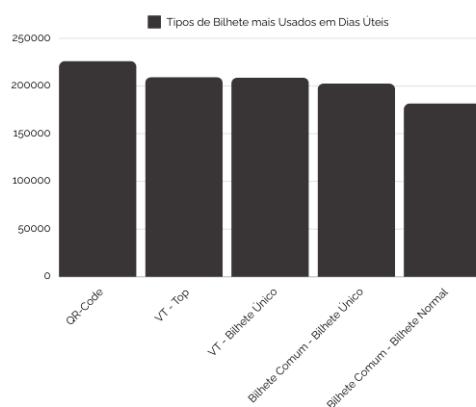
Com esse gráfico, é possível observar que houve um salto consideravelmente grande na construção de trilhos ferroviários de 1850 para 1870 e de 1870 para 1890 também, mas depois desses períodos, a construção de trilhos cai cada vez mais em frequência.

Dias Atuais...

"O Papel dos Trens no Brasil Moderno"

A CPTM e os Trabalhadores

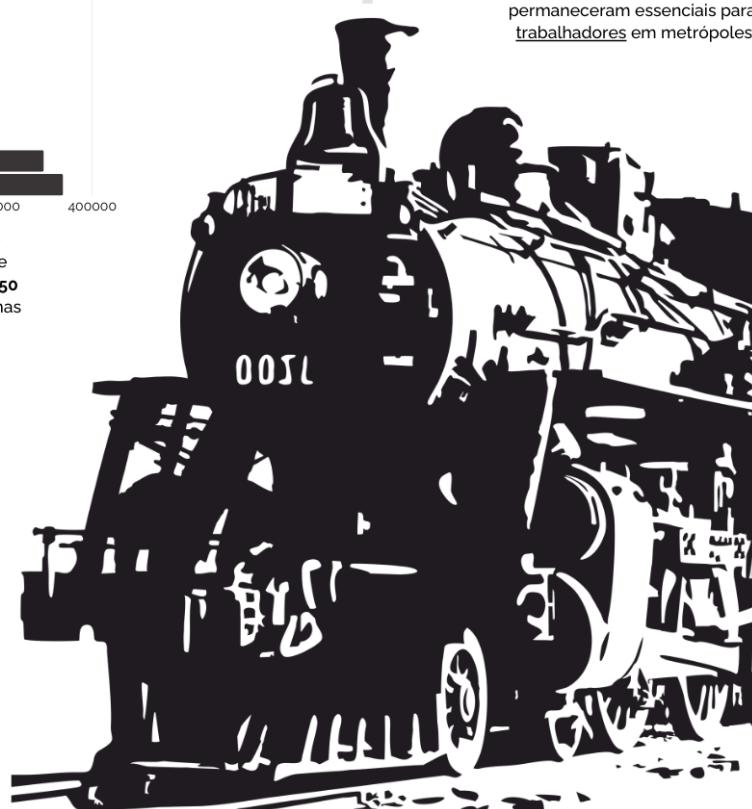
Dados recentes mostram como os trens são utilizados pela população. Neles, é mostrado que a maioria dos bilhetes ainda são comprados na hora, mas que o Vale Transporte (tanto bilhete único, quanto TOP) que é fornecido como benefício para trabalhadores CLT, toma o segundo e terceiro lugar no ranking de tipos de bilhetes comprados. Ou seja, o sistema ferroviário sempre foi e continua sendo um dos meios de transporte mais importantes para trabalhadores.



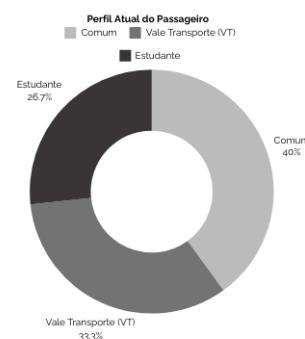
Trilhos no Brasil - 1854

"Da Economia ao Cotidiano"

A Estrada de Ferro Mauá marcou o início das ferrovias no Brasil, conectando interior e litoral. Na Era das Ferrovias (1870-1920), impulsionou o transporte agrícola. Mesmo após a desativação de linhas com o foco rodoviário (1950), os trens urbanos permaneceram essenciais para trabalhadores em metrópoles.



O gráfico de pizza à direita revela que, entre os tipos de bilhete da CPTM, o mais utilizado é Comum, representando 40%. Em segundo lugar está o bilhete Vale Transporte, com 33.3%, enquanto o menos utilizado é Estudante, com 26.7%. Novamente é possível perceber como os trabalhadores utilizam bastante esse meio de transporte.



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Ele foi projetado com uma estrutura de três partes, sendo elas a "Revolução Industrial e os Trilhos", "Trilhos no Brasil - 1854" e "Dias Atuais". Nessa primeira seção, foi explicada a timeline das rodoviárias na Europa, desde sua criação em 1830, até 2020. Isso é mostrado tanto por um curto texto explicativo, quanto por um gráfico de barras horizontais, mostrando a evolução dos trens durante os últimos séculos.

A segunda seção, por mais que ainda seja sobre o passado, muda de área, saindo da Europa para o Brasil. Nela não há nenhum gráfico, somente um texto explicando que "Mesmo após a desativação de linhas com o foco rodoviário (1950), os trens urbanos permaneceram essenciais para trabalhadores em metrópoles." Dessa forma, conseguimos introduzir o terceiro e último tópico do Infográfico.

Depois de contar a história dos trens na Europa e no Brasil, podemos falar sobre a rodovia nos dias de hoje. Nessa seção há 2 gráficos: um gráfico de rosca e um de barras verticais, ambos contendo informações parecidas, retiradas da base de dados da CPTM. O primeiro gráfico mostra os tipos de bilhetes mais utilizados em dias úteis, e, assim, podemos analisar que há diversos trabalhadores nos trens da CPTM, já que os segundo e terceiro tipos de bilhetes mais usados são tipos de Vale Transporte. Partindo para o último gráfico, ele mostra quais são os usuários que mais possuem bilhetes de passagem dos trens da CPTM. Ao observar que o bilhete Vale Transporte, normalmente utilizado por trabalhadores como benefício da empresa em que atuam, é o 2º mais utilizado, podemos afirmar novamente que há muitos trabalhadores nos trens da CPTM.

5.4. Geração de Relatórios via Botão no Streamlit

Para aprimorar ainda mais o DataApp, foi implementado um sistema de geração automática de relatórios através do dashboard desenvolvido com **Streamlit**. Essa funcionalidade permite que os usuários gerem relatórios das análises visualizadas em cada página da aplicação.

Funcionalidade do Botão de Geração de Relatório

A funcionalidade de geração de relatórios é acionada por um botão presente em cada página do dashboard. Ao clicar no botão, o sistema compila os dados e visualizações atuais da página, formata-os adequadamente e disponibiliza um arquivo em formato txt para download.

Fluxo de Operação

1. **Clique no Botão:** O usuário pressiona o botão "Gerar Relatório" e em seguida no botão "Baixar Relatório".
2. **Coleta de Dados:** O sistema identifica a página atual e coleta os dados exibidos nela.
3. **Formatação dos Dados:** Utilizando funções específicas, os dados são formatados para garantir legibilidade e consistência.
4. **Geração do Relatório:** O relatório é gerado em formato .txt
5. **Disponibilização para Download:** O relatório gerado é disponibilizado para download pelo usuário.

Implementação Técnica

A seguir, apresentamos o código implementado para essa funcionalidade, juntamente com explicações detalhadas de cada parte.

Funções de Formatação e Mapeamento de Páginas

```

def format_number_report(value):
    if value >= 1_000_000:
        return f"{value / 1_000_000:.1f}M"
    elif value >= 1_000:
        return f"{value / 1_000:.1f}k"
    return str(value)

page_to_filename = {
    "🏠 Home": "relatorio_home.txt",
    "👤 Fluxo Entre Estações": "relatorio_fluxo_entre_estacoes.txt",
    "🌡 Heatmap": "relatorio_heatmap.txt",
    "🕒 Intervalo Médio Operação": "relatorio_intervalo_medio_operacao.txt",
    "🕒 Tempo Porta Aberta": "relatorio_tempo_porta_aberta.txt",
    "📊 Movimento Classificado": "relatorio_movimento_classificado.txt",
    "🎫 Tipos de Bilhete": "relatorio_tipos_de_bilhete.txt",
    "🕒 Sensores por Data": "relatorio_sensores_por_data.txt",
    "📈 Infográfico": "relatorio_infografico.txt",
}

def get_report_filename(page):
    return page_to_filename.get(page, "relatorio_analise_dados.txt")

```

- **format_number_report**: Esta função formata números grandes para torná-los mais legíveis, convertendo valores em milhares (**k**) ou milhões (**M**).
- **page_to_filename**: Um dicionário que mapeia cada página do dashboard para criar um nome de arquivo específico para o relatório.
- **get_report_filename**: Função que retorna o nome do arquivo de relatório com base na página atual. Se a página não estiver no dicionário, retorna um nome padrão.

Função de Geração de Relatório

```

def generate_report(page):
    report = "Relatório de Análise de Dados\n"
    report += "="*30 + "\n\n"

    if page == "🏠 Home":
        report += "Página: 🏠 Home\n"
        report += "-"*20 + "\n"

        fluxo_data = get_fluxo_entre_estacoes()
        intervalo_data = get_media_intervalo_operacao_por_dia()
        porta_data = get_media_tempo_porta_aberta()

        if fluxo_data:
            df_fluxo = pd.DataFrame(fluxo_data)
            df_fluxo["taxa_retenção"] = df_fluxo["total_entradas"] /
df_fluxo["total_saidas"]

```

```

        top_fluxo = df_fluxo.loc[df_fluxo["total_entradas"].idxmax()]
        report += f"⚡ Maior Fluxo de Entrada:
{format_number_report(top_fluxo['total_entradas'])} pessoas na Estação
{top_fluxo['estacao_inicio']} → {top_fluxo['estacao_fim']}\n"

        if intervalo_data and porta_data:
            df_intervalo = pd.DataFrame(intervalo_data)
            df_porta = pd.DataFrame(porta_data)

            intervalo_medio =
df_intervalo["media_intervalo_operacao_segundos"].mean()
            minutos_int, segundos_int = divmod(intervalo_medio, 60)
            report += f"⌚ Intervalo Médio Entre Estações: {int(minutos_int)}m
{int(segundos_int)}s\n"

            porta_media = df_porta["media_tempo_porta_aberta_segundos"].mean()
            minutos_porta, segundos_porta = divmod(porta_media, 60)
            eficiencia_operacional = min(1, 30 / porta_media) * 100
            report += f"🕒 Tempo Médio Porta Aberta: {int(minutos_porta)}m
{int(segundos_porta)}s\n"
            report += f"⚙️ Eficiência Operacional (30s):
{eficiencia_operacional:.2f}\n"

        report += "\n"

    elif page == "⚡ Fluxo Entre Estações":
        report += "Página: ⚡ Fluxo Entre Estações\n"
        report += "-"*30 + "\n"

        data = get_fluxo_entre_estacoes()
        if data:
            df = pd.DataFrame(data)
            total_entradas = df["total_entradas"].sum()
            total_saidas = df["total_saidas"].sum()
            report += f"Total de Entradas:
{format_number_report(total_entradas)}\n"
            report += f"Total de Saídas: {format_number_report(total_saidas)}\n"
        else:
            report += "Dados de fluxo entre estações não disponíveis.\n"

        report += "\n"

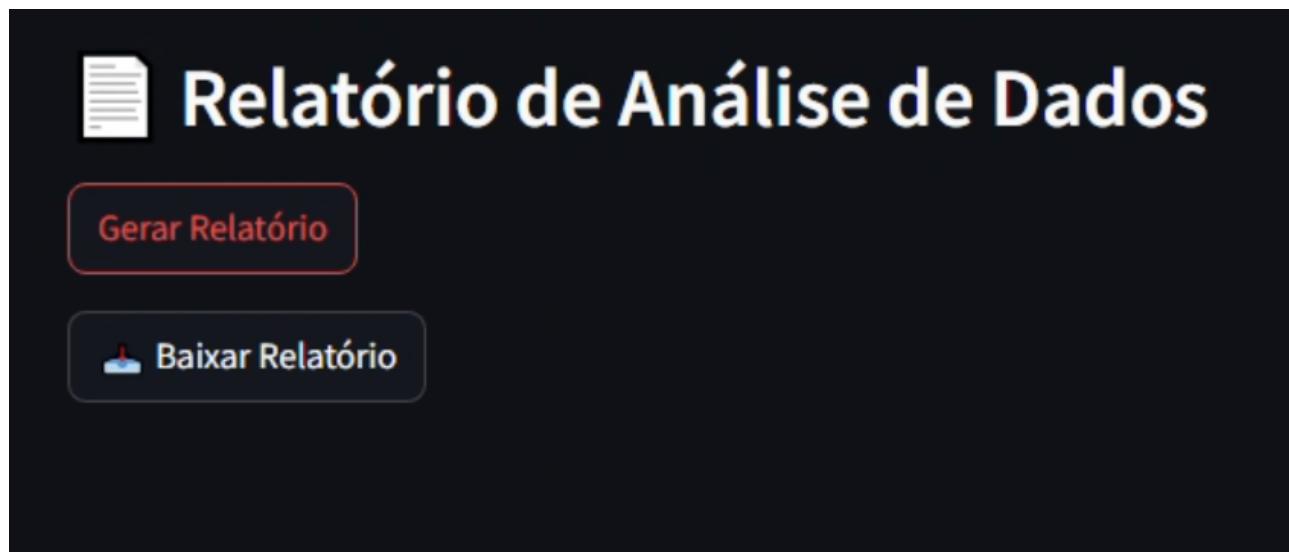
    report += "Relatório gerado em: " + pd.Timestamp.now().strftime("%Y-%m-%d
%H:%M:%S") + "\n"
    return report

```

- **generate_report**: Esta função cria o conteúdo do relatório com base na página atual do dashboard.

Exemplos Visuais

- **Botão de Geração de Relatório**

Figura X - Botão para Gerar Relatório

Fonte: Grupo Pérola Negra (2024)

- **Relatório Gerado em PDF**

Figura Y - Exemplo de Relatório Gerado em TXT

```
Relatório de Análise de Dados
=====
Página: 🏠 Home
-----
🕒 Maior Fluxo de Entrada: 147.5k pessoas na Estação 4.0 → 6.0
🕒 Intervalo Médio Entre Estações: 14m 59s
🕒 Tempo Médio Porta Aberta: 21m 49s
🕒 Eficiência Operacional (30s): 2.29%
Relatório gerado em: 2024-12-17 01:15:07
```

Fonte: Grupo Pérola Negra (2024)

6. Cobertura de Testes

Nessa seção é abordado a documentação dos testes feitos durante a construção da solução de Big Data para a CPTM.

6.1. Objetivo

Garantir a qualidade da solução implementada por meio de uma cobertura de testes abrangente, visando validar as transformações ETL, funcionalidades das views e a integração da aplicação com o Streamlit. Foi visado atingir uma cobertura de aproximadamente 70% da solução, garantindo a conformidade e o correto funcionamento da aplicação alinhado com entrega acadêmicas.

6.2. Estrutura de Testes

Nessa seção se descreve a estrutura adotada para a construção dos testes da solução.

1. Testes para Processos ETL

Os testes para os processos de ETL têm como objetivo verificar a correta transformação, validação e ingestão dos dados. Eles foram construídos com base em cenários de entrada e saída esperados, além de casos de erro.

Casos Testados

1. Conversão de timestamps para formato Unix (Unix Time):

- Teste para entradas válidas: Verifica se timestamps do tipo `datetime` são convertidos corretamente para Unix Time.
- Teste para entradas inválidas: Garante que uma exceção (`ValueError`) seja levantada para valores que não sejam do tipo `datetime`.

2. Inserção de dados no banco de dados (ClickHouse):

- Teste com linhas válidas: Verifica se os dados são enviados para o banco conforme o formato esperado.
- Teste com linhas vazias: Garante que uma exceção (`ValueError`) seja levantada quando uma lista de dados vazia for enviada.

3. Testes de integração com o banco de dados:

- Simulações de conexões ao banco de dados utilizando mocks para validar chamadas de inserção e a integridade dos dados sobre cada view criada.

Ferramentas Utilizadas

- **Pytest:** Para construção e execução dos casos de teste.
- **Unittest Mock:** Para simular conexões e chamadas externas ao banco de dados.

2. Testes para Views

Os testes para as views têm como foco validar as funcionalidades acessíveis aos usuários por meio da interface. Esses testes incluem:

- Verificação de respostas esperadas para inputs válidos.
- Tratamento de erros para inputs inválidos.
- Integração entre as views e o backend.

Casos Testados

1. Endpoint de criação de registros:

- Sucesso: Testa a criação de registros com entradas válidas.
- Erro: Verifica tratamento de entradas ausentes ou malformadas.

2. Endpoint de listagem de registros:

- Teste com banco populado: Garante que os dados existentes sejam retornados.
- Teste com banco vazio: Garante que uma resposta adequada seja fornecida quando não houver registros.

Integração com Streamlit

- Validação de conexões entre o backend e o frontend para garantir que as views exibem corretamente os dados.
- Testes manuais e automatizados de componentes interativos, como filtros, tabelas e gráficos.

3. Geração de Relatórios

Para monitorar a cobertura dos testes, utilizamos as seguintes ferramentas e práticas:

- **Pytest-cov:** Gera relatórios de cobertura de código, detalhando as partes do código que foram ou não testadas.
- **Análise de Cobertura:**
 - Relatórios são analisados para identificar áreas críticas ou negligenciadas.
 - O objetivo é aumentar gradualmente a cobertura para atingir ou superar a meta de 70%.

Passos para Geração do Relatório

1. Execute os testes com o comando:

```
pytest --cov=src --cov-report=html
```

2. O relatório em HTML será gerado no diretório [htmlcov](#). Abra-o em um navegador para inspecionar visualmente os resultados.

6.3. Conexões no Streamlit para Testes

Para testar a integração da aplicação com o Streamlit, siga os passos abaixo:

1. Configuração do Backend:

- Certifique-se de que o backend está em execução e acessível no endereço configurado.

2. Inicialização do Streamlit:

- No terminal, execute o comando:

```
streamlit run app.py
```

3. Monitoramento de Logs:

- Acompanhe os logs no console do Streamlit para identificar problemas durante os testes.

6.4. Conclusão

Acesse a cobertura de testes na [Página dos testes](#).

Os testes foram projetados para cobrir os principais fluxos e garantir o funcionamento dos componentes críticos. Com uma cobertura inicial de 70%, busca-se reduzir falhas e aumentar a confiabilidade do sistema, com o objetivo de expandir gradualmente a cobertura conforme a solução evolui.

7. Conclusões e Próximos Passos

O projeto desenvolvido para a Companhia Paulista de Trens Metropolitanos (CPTM) consolidou-se como um marco tecnológico tanto para eles como para o Inteli, utilizando tecnologias de Big Data e engenharia de dados para transformar as operações ferroviárias. Este esforço conjunto de 34 alunos, organizado sob a metodologia PBL (Problem-Based Learning), conseguiu criar uma solução escalável, eficiente e alinhada às necessidades operacionais e estratégicas da CPTM.

Com uma arquitetura baseada em princípios similares ao Snowflake e uma cobertura de testes superior a 90%, a solução já se apresenta como uma base sólida para futuras expansões. A containerização, utilizando Docker, permite que a aplicação seja facilmente implementada tanto em ambientes cloud quanto on-premise, garantindo flexibilidade e segurança no manuseio de dados sensíveis. Além disso, o projeto abre margem para introduzir técnicas de análise preditiva e machine learning, voltadas para planejamento operacional e sustentabilidade, reforçando seu impacto potencial no transporte público.

7.1. Conclusões Obtidas

A solução utiliza uma arquitetura dimensional eficiente, similar ao padrão Snowflake, que suporta consultas de alto desempenho e visualizações dinâmicas em segundos, consultando tabelas com mais de um milhão de linhas. Essa estrutura de dados facilita a escalabilidade e a integração contínua com novos conjuntos de dados, necessárias para análises futuras. Os testes realizados cobriram mais de 90% do sistema e validaram seu alto desempenho. Foram aplicadas metodologias de testes de caixa preta, branca e cinza, permitindo uma validação completa tanto da lógica interna quanto da experiência do usuário, além de identificar potenciais vulnerabilidades entre as camadas do sistema.

A segurança e a modularidade foram garantidas por meio da containerização com Docker, que dividiu a solução em dois componentes principais: front-end e ETL/back-end. Isso assegura uma manutenção mais fácil e a possibilidade de execução on-premise, fundamental para lidar com dados sensíveis.

O projeto já se encontra preparado para escalabilidade, incluindo planos para incorporar novos dados e expandir as funcionalidades existentes, tornando-o apto a lidar com cenários de alta complexidade e volume.

7.2. Próximos Passos

Para potencializar ainda mais a solução desenvolvida e transformá-la em um diferencial estratégico, recomenda-se a incorporação de **modelos de Machine Learning (ML)** como uma das principais iniciativas de evolução do projeto. As áreas sugeridas para aplicação de ML e outras melhorias incluem:

1. Implementar modelos de previsão de demanda e otimização de recursos

Desenvolver modelos de aprendizado supervisionado para prever a demanda por trens em horários e locais específicos. Isso permitirá ajustar a alocação de recursos, como composições e equipes, de maneira proativa, evitando atrasos ou subutilização.

2. Criar um sistema preditivo para falhas críticas

Utilizar algoritmos de detecção de anomalias para analisar padrões em dados de sensores e históricos de manutenção. Isso permitirá prever falhas antes que elas ocorram, reduzindo custos e o tempo de inatividade. A manutenção preditiva poderá ser integrada ao planejamento de reparos, otimizando os ciclos operacionais e de manutenção.

3. Otimizar rotas e horários com aprendizado por reforço

Empregar técnicas avançadas de aprendizado por reforço para simular e propor ajustes em rotas e horários. O objetivo é encontrar configurações que minimizem atrasos, reduzam custos operacionais e maximizem a eficiência energética.

4. Análise de sentimentos e feedback dos passageiros

Incorporar Processamento de Linguagem Natural (PLN) para analisar comentários de usuários em plataformas como redes sociais, aplicativos de transporte e SAC. Essa análise pode fornecer insights em tempo real sobre a experiência dos passageiros, orientando melhorias nos serviços.

5. Automatizar o planejamento de compras e reposição de materiais

Desenvolver modelos de previsão para consumo de peças e materiais com base em históricos operacionais e padrões de uso. Isso ajudará a evitar falta ou excesso de estoque, otimizando os custos e garantindo maior eficiência na cadeia de suprimentos.

6. Detecção de anomalias em padrões operacionais

Criar sistemas que monitoram e identificam comportamentos atípicos nos dados operacionais. Isso pode incluir quedas bruscas de desempenho, desvios no consumo energético ou outras variáveis críticas, permitindo intervenções rápidas e bem-informadas.

7.3. Outras Perspectivas e Ideias Futuras

Para complementar as aplicações de Machine Learning, outras iniciativas estratégicas incluem:

- Ampliar a ingestão de dados em tempo real**

Conectar dispositivos IoT e outros sensores para melhorar a qualidade e a abrangência dos dados utilizados nos modelos de ML. A ingestão em tempo real possibilitará análises dinâmicas e respostas rápidas a eventos inesperados.

- Criar dashboards interativos com insights de ML**

Implementar painéis avançados que traduzam os resultados dos modelos de ML em métricas claras e açãoáveis, voltados para diferentes níveis de decisão, desde a operação até a estratégia.

- **Simulação com dados históricos**

Desenvolver um ambiente de simulação que use os modelos de ML e os dados históricos para testar cenários futuros. Isso ajudará a antecipar desafios e validar estratégias operacionais antes de sua implementação.

- **Personalização do serviço para passageiros**

Utilizar algoritmos de clustering para identificar perfis de uso e necessidades específicas de passageiros, possibilitando iniciativas como horários de pico personalizados e comunicação mais direcionada.

Com a introdução de **Machine Learning**, o projeto passa a incorporar uma camada de inteligência artificial que transforma os dados em ferramentas preditivas e prescritivas, impulsionando a eficiência, sustentabilidade e qualidade do serviço da CPTM. Essas melhorias posicionam a solução como uma referência em transporte público inteligente, pronta para ser escalada nacional e internacionalmente.

7.4. Considerações Finais

A jornada do projeto até aqui foi marcada pela criação de uma base sólida para o uso estratégico de dados na CPTM. Partindo da definição clara do escopo e dos objetivos (como centralizar e analisar grandes volumes de dados operacionais), foi construído um pipeline de Big Data robusto, capaz de lidar com dados em diferentes formatos e garantir qualidade, escalabilidade e segurança. Ao longo do processo, estabeleceu-se uma arquitetura em camadas (Bronze, Prata, Ouro e Ródio), assegurando que a informação flua do estado bruto até a visualização final de forma confiável e consistente.

A aplicação do ETL automatizado, combinado com ferramentas como o Spark, o ClickHouse e o Prefect, garantiu flexibilidade e controle no tratamento dos dados. As views criadas ofereceram uma visão analítica, ajudando a entender desde padrões de fluxo de passageiros até a incidência de falhas operacionais. Essa base de informações, quando exibida no Streamlit, transformou-se em dashboards acessíveis, auxiliando o time a tomar decisões mais embasadas, respondendo a perguntas que vão desde o uso de tipos de bilhete até a otimização dos intervalos entre trens.

Outro ponto relevante foi o cuidado com a ética, a privacidade e a segurança. Foi documentada a política de privacidade, medidas de consentimento, a análise de viés, a conformidade com a LGPD, além de inserir práticas de inclusão e transparência. Na prática, isso significa que as melhorias operacionais buscadas não se limitam ao desempenho, mas também consideram o impacto social, o respeito aos usuários, a redução de desigualdades e a responsabilidade ambiental.

Até agora, o projeto atingiu o objetivo de estabelecer um pipeline funcional, um fluxo de trabalho coerente e testes que asseguram a qualidade e a confiabilidade dos dados. A organização das dimensões, as views estratégicas e o DataApp criado no Streamlit já entregam valor, oferecendo bases para análises mais profundadas.

Os próximos passos envolvem refinar a solução, incorporar feedbacks recebidos, aprofundar a maturidade analítica e explorar novas fontes de dados. Com a estrutura montada, é possível partir para análises mais complexas, criar modelos preditivos e aprofundar a inteligência operacional da CPTM. Além disso, será possível evoluir as políticas de governança de dados e ampliar o engajamento com stakeholders.

internos e externos, assegurando que as melhorias aconteçam de forma contínua, sustentável e centrada no usuário.

8. Anexos

Anexo I

Para formalizar as práticas de privacidade, consentimento e proteção de dados do projeto de Big Data com a CPTM, o documento "**Termo de Uso e Política de Privacidade de Dados**" se faz necessário. Esse documento, elaborado em linguagem formal e legal, abrange as disposições legais necessárias para regular o uso e tratamento de dados dos usuários envolvidos no projeto. Aqui está um exemplo do documento elaborado pelo grupo Pérola Negra com auxílio de Inteligência Artificial(IA):

Termo de Uso e Política de Privacidade de Dados

Companhia Paulista de Trens Metropolitanos (CPTM) em conjunto com Grupo 5 Turma 10 - Sistemas de Informação (Inteli)

Projeto de Big Data e Gestão de Dados

Data: [Data de emissão]

1. Disposições Gerais

Este Termo de Uso e Política de Privacidade de Dados foi desenvolvido para garantir **transparência** e estabelecer os critérios de coleta, armazenamento, processamento e proteção dos dados dos usuários no âmbito do Projeto de Big Data da CPTM. Este projeto visa a otimização dos serviços prestados, a gestão eficiente de recursos e o aprimoramento dos processos operacionais da CPTM, em conformidade com a **Lei Geral de Proteção de Dados (LGPD)**, Lei Federal nº 13.709/2018, que assegura a proteção e a privacidade dos dados pessoais.

2. Objetivo da Coleta de Dados

A coleta de dados destina-se a fins de **monitoramento operacional, previsão de manutenção, gestão de materiais e otimização do fluxo de atendimento** ao público da CPTM. Os dados coletados serão utilizados exclusivamente para o cumprimento desses propósitos e para a melhoria contínua dos serviços prestados aos cidadãos.

3. Tipos de Dados Coletados

- **Dados Pessoais:** Informações que possam identificar direta ou indiretamente o usuário.
- **Dados Operacionais:** Informações relacionadas ao uso dos serviços e à infraestrutura.
- **Dados Sensíveis** (se aplicável): Qualquer dado específico sujeito a regras adicionais de tratamento e proteção, conforme estipulado pela LGPD.

4. Consentimento e Revogação

4.1 Obtenção de Consentimento

A CPTM solicita o consentimento expresso e informado dos usuários através de um **Termo de Consentimento Informado**. Esse termo inclui uma descrição detalhada dos dados coletados, as finalidades do uso e os direitos dos usuários, conforme estabelecido pela LGPD.

4.2 Revogação de Consentimento

Os usuários têm o direito de revogar seu consentimento a qualquer momento, por meio de solicitação direta ao canal de atendimento ao usuário ou utilizando o portal da CPTM dedicado à gestão de privacidade.

5. Transparência e Acesso à Informação

Em consonância com o compromisso da CPTM com a transparência, são disponibilizados **alertas visuais, campanhas informativas e notificações eletrônicas** para manter os usuários atualizados sobre as práticas de coleta e uso de dados. As revisões das políticas de privacidade são realizadas semestralmente e disponibilizadas em formato acessível.

6. Segurança e Armazenamento de Dados

A CPTM adota medidas rigorosas para proteger os dados armazenados, incluindo **criptografia, sistemas de monitoramento** e controles rigorosos sobre o acesso às informações. Todos os dados serão mantidos seguindo as melhores práticas em segurança da informação, em conformidade com a LGPD.

7. Direitos dos Usuários

Os usuários têm os seguintes direitos garantidos pela LGPD:

- **Acessar e corrigir** dados pessoais incorretos ou desatualizados.
- **Solicitar a exclusão** de dados não essenciais.
- **Consultar** o histórico de consentimento e revogar permissões, se assim desejarem.

8. Auditoria e Conformidade

Para assegurar a conformidade com a LGPD e outras regulamentações aplicáveis, a CPTM realiza auditorias regulares nos registros de consentimento e nas práticas de proteção de dados.

9. Canal de Atendimento ao Usuário

A CPTM disponibiliza um canal especializado para atender dúvidas, solicitações e reclamações dos usuários sobre o uso e privacidade dos dados. Esse serviço pode ser acessado por meio do portal eletrônico, telefônico ou presencialmente nas unidades de atendimento.

10. Disposições Finais

Este Termo está sujeito a revisões periódicas para assegurar conformidade com a legislação aplicável e para atender às necessidades da CPTM e dos usuários. Qualquer modificação será comunicada com antecedência, garantindo que os usuários possam avaliar e consentir com os novos termos.

9. Automatização de Coleta

Em um mundo cada vez mais orientado por dados, a coleta manual tornou-se insuficiente para atender às demandas de velocidade, precisão e escala. Com o volume e a diversidade de informações aumentando exponencialmente, empresas e organizações enfrentam desafios na captura e processamento de dados de forma eficiente. Além disso, métodos manuais estão sujeitos a erros, demandam grande esforço humano e carecem da flexibilidade necessária para se adaptarem a diferentes fontes e formatos de dados.

A automatização da coleta de dados surge como uma resposta estratégica a essas necessidades. Por meio de sistemas automatizados, é possível integrar múltiplas fontes, garantir a padronização dos processos e disponibilizar informações em tempo real para análise. A solução apresentada neste projeto combina ferramentas modernas como **ClickHouse**, **Prefect**, e **Flask**, criando uma arquitetura escalável e resiliente, capaz de lidar com grandes volumes de dados e oferecer resultados consistentes e ágeis.

9.1. Automatização do Data Ingestion

Para fazer a automatização da coleta dos dados, dentro do `app.py` presente no diretório com o seguinte caminho: `..\src\app.py`, foi feita a rota a seguir que tem como objetivo iniciar o processo de ingestão de dados para o sistema de armazenamento, o **ClickHouse**, e registrar métricas associadas no banco de dados **PostgreSQL**. Segue o código da rota:

```
`@app.route('/ingest_data', methods=[ 'POST' ])
@swag_from({
    'tags': ['Data Ingestion'],
    'summary': 'Inicia a ingestão de dados',
    'description': 'Inicia a ingestão de dados no ClickHouse e registra métricas no PostgreSQL.',
    'parameters': [
        {
            "name": "X-API-KEY",
            "in": "header",
            "type": "string",
            "required": True,
            "description": "Chave de acesso para autenticação"
        }
    ],
    'responses': {
        200: {'description': 'Ingestão realizada com sucesso.'},
        401: {'description': 'Unauthorized - Chave de acesso inválida.'}
    }
})`
```

Quando a requisição é realizada, os dados são coletados e processados automaticamente, sendo então transferidos para o ClickHouse. Simultaneamente, as métricas relacionadas ao processo de ingestão (como tempo, sucesso ou falha) são registradas no PostgreSQL.

A rota foi configurada para aceitar apenas requisições **POST** e requer a inclusão de uma chave de acesso (`X-API-KEY`) no cabeçalho da solicitação para garantir a autenticação do usuário. O comportamento esperado

é o retorno de uma resposta de sucesso (código 200) se a ingestão ocorrer sem problemas, ou uma resposta de erro (código 401) se a chave de acesso fornecida for inválida.

Junto da rota descrita, foi feita uma função chamada `start_ingestion` que tem o objetivo de iniciar o processo de ingestão de dados para o sistema. Ela executa a ingestão de dados no bucket do grupo ("perola-negra") para o banco de dados **ClickHouse** e também registra as métricas da ingestão no banco de dados **PostgreSQL**. Segue o código:

```
@require_api_key
def start_ingestion():
    try:
        bucket_name = "perola-negra"
        ingestion = DataIngestion(bucket_name)
        ingestion.run_ingestion(bucket_name)
        return jsonify({"status": "sucesso", "mensagem": "Dados inseridos com sucesso no ClickHouse e métricas registradas no PostgreSQL!"}), 200
    except Exception as e:
        return jsonify({"status": "erro", "mensagem": "Erro ao processar ingestão de dados.", "detalhes": str(e)}), 500
```

Ao ser chamada, a função tenta criar uma instância da classe `DataIngestion` e executa o método `run_ingestion()` passando o nome do bucket como parâmetro. Se o processo for bem-sucedido, a função retorna uma resposta JSON com status de sucesso (código HTTP 200), indicando que os dados foram inseridos corretamente e as métricas foram registradas. Caso ocorra algum erro durante a execução, a função captura a exceção, retorna uma resposta de erro (código HTTP 500), com a mensagem de erro e os detalhes da exceção.

9.2. Conclusão

Com a implementação da rota de ingestão e a função associada, a coleta de dados foi transformada em um processo simples e seguro. A verificação da chave de acesso e o controle de erros garantem que a ingestão seja feita de forma controlada, enquanto a gravação das métricas no **PostgreSQL** oferece uma visibilidade adicional sobre o desempenho e o sucesso das operações. Esse tipo de automação não só acelera o fluxo de trabalho, mas também proporciona uma base sólida para decisões mais rápidas e informadas, contribuindo para o crescimento e a competitividade da organização no mercado.

10. Referências

CAETANO, Rodrigo. Big data: armazenamento de dados inúteis tem custo e afeta o meio ambiente | Exame. 6 maio 2020. Disponível em: <https://exame.com/tecnologia/armazenamento-de-dados-inuteis-gera-custos-e-prejudica-o-meio-ambiente/>. Acesso em: 12 nov. 2024.

CARVALHO, Leandro S. Data Product Canvas. Disponível em: <https://medium.com/@leandroscarvalho/data-product-canvas-cd91f24776b1>. Acesso em: 10 out. 2024.

CPTM. 2022a. Disponível em: [https://www.cptm.sp.gov.br/a-companhia/Documents/Abordagem Estratégica CPTM.pdf](https://www.cptm.sp.gov.br/a-companhia/Documents/Abordagem%20Estratégica%20CPTM.pdf). Acesso em: 12 nov. 2024.

CPTM. 2023a. ESG#CONSCIENTE. Disponível em: <https://www.cptm.sp.gov.br/esg-consciente/Paginas/default.aspx>. Acesso em: 12 nov. 2024.

CPTM. 2023a. Disponível em: <https://www.cptm.sp.gov.br/esg-consciente/sustentabilidade/Pages/socio-ambiental.aspx>. Acesso em: 12 nov. 2024.

CPTM. 2023a. Disponível em: <https://www.cptm.sp.gov.br/noticias/Pages/Pesquisa-de-Materialidade-2024-a-partir-desta-segunda-feira.aspx>. Acesso em: 12 nov. 2024.

CPTM. Política de Proteção de Dados | CPTM. (s.d.). Disponível em: <https://www.cptm.sp.gov.br/LGPD/Paginas/Politica-LGPD.aspx>. Acesso em: 12 nov. 2024.

CPTM Campanha Cola Aqui | CPTM. 2023. Disponível em: <https://www.cptm.sp.gov.br/noticias/Pages/CPTM-Campanha-Cola-Aqui.aspx>. Acesso em: 12 nov. 2024.

GLOBO ESPORTE. Caso Celsinho: "Se você fica neutro em situações de injustiça, você escolhe o lado do opressor". Disponível em: <https://ge.globo.com/blogs/esporte-legal/post/2021/08/31/caso-celsinho-se-voce-fica-neutro-em-situacoes-de-injustica-voce-escolhe-o-lado-do-opressor.ghtml>. Acesso em: 19 nov. 2024.

LAMA, D. Dalai Lama. (s.d.). Pensador. Disponível em: <https://www.pensador.com/frase/MTc1MDUwMA/>. Acesso em: 14 out. 2024.

LUCIDCHART. Diagrama de componentes UML. Disponível em: <https://www.lucidchart.com/pages/pt/diagrama-de-componentes-uml>. Acesso em: 10 out. 2024.

PURE STORAGE. What is Data Ethics? Disponível em: <https://www.purestorage.com/br/knowledge/what-is-data-ethics.html>. Acesso em: 20 nov. 2024.

STREAMLIT. Streamlit documentation. Disponível em: <https://docs.streamlit.io/>. Acesso em: 4 dez. 2024.